Distribution Agreement

Signature:

Yanting Huang   Date

Disease Risk annotation of Genomic and Epigenomic Variants using Machine Learning Approaches

By

Yanting Huang

Doctor of Philosophy

Computer Science

____

Zhaohui Qin, Ph.D.

Advisor

____

Hao Wu, Ph.D.

Committee Member

____

Peng Jin, Ph.D.

Committee Member

____

Matthew Reyna, Ph.D.

Committee Member

Accepted: ____

Kimberly Jacob Arriola, Ph.D.

Dean of the James T. Laney School of Graduate Studies

____

Date

Disease Risk annotation of Genomic and Epigenomic Variants using Machine Learning Approaches

By

Yanting Huang
B.S., Southeast University, 2016

Advisor: Zhaohui Qin, 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
2022

Abstract

Disease Risk annotation of Genomic and Epigenomic Variants using Machine Learning Approaches
By Yanting Huang

There has been a tremendous quantity of omics data produced by high-throughput genomics technologies nowadays. Understanding the impact of genomic variations and epigenomic modifications is important for discovering the mechanism of complex diseases. Over the last two decades, thousands of genome-wide association studies (GWASs) and epigenome-wide association studies (EWASs) have identified tens of thousands of disease-susceptibility loci that are associated with certain diseases. In addition to the association studies, current progress of machine learning and deep learning studies have pushed the edge and provided great opportunities to integrate omics data to uncover complicated relationships of features from different aspects of regulatory factors for the disease risk annotations of genomic and epigenomic variants. By utilizing comprehensive omics data from the The Encyclopedia of DNA Elements (ENCODE) and the Roadmap Epigenomics Mapping Consortium (REMC) projects, I proposed several machine learning predictive models with different focuses on genomic and epigenomic variants annotations, which includes 1) EWASplus, an ensemble learning based framework for the risk prediction of DNA methylation loci associated with Alzheimer's Disease, 2) CASAVA (Disease Category-specific Annotation of Variants), a disease category specific risk annotation for the whole genome wide SNPs (single nucleotide polymorphism), 3) DRAFT (Disease Risk Annotation with Few shoTs learning), an end-to-end deep learning based approach that incorporates contrastive learning to tackle the lack of risk variants that hinder the application of traditional deep learning models to this research field.

Disease Risk annotation of Genomic and Epigenomic Variants using Machine Learning Approaches

By

Yanting Huang
B.S., Southeast University, 2016

Advisor: Zhaohui Qin, 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
2022

Acknowledgements

First and foremost, I would like to thank my PhD advisor, Dr. Steve Qin for his selfless guidance and help throughout the past six years. He is my go-to person whenever I encounter challenges or receive good news. As a result of his passion for academics and trustworthiness in life, he has shed light on my life while at Emory. I will remember the time we spent discussing project ideas, the time we spent working on the manuscript and revisions, and the help you provided me when I was searching for a job. I have had the great honor to be your student, and you are my lifetime role model in every aspect of my life.

Next, I would like to thank all the senior PIs and professors who have helped me during my PhD for their thoughtful comments and considerate guidance, Dr. Peng Jin, Dr. Hao Wu, Dr. Matthew Reyna, Dr. Thomas Wingo, and Dr. Tianwei Yu. I have gained a great deal of experience in the field of research as a result of your valuable comments and insightful suggestions. Furthermore, I would like to thank all of my collaborators, especially Dr. Shihua Zhang and Dr. Zhen Cao, who had been working with me on the CASAVA project. I will remember the helpful meeting we had, the inspiring communications we had and how we overcame the time lag between the US and China until the project was successfully completed.

Also, I would like to express gratitude to all my friends at Emory. I am extremely lucky to meet you, get to know you and be able to accompany you on this journey through life. It is impossible for me to make this journey without your continued support. My sincere thanks go to Huan He, Jing Zhang, Qingyang Zhu, Shaojun Yu, Ziwei Dong, Hongwen Song, Wenkai Zheng, Wenjing Ma, Jiaying Lu and Feng Zhu among others.

Finally, I must express my very profound gratitude to my parents and to my girlfriend, Tianyu Lan, for providing me with love, unfailing support and continuous encouragement throughout my years of study. This accomplishment would not have been possible without them.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

In recent years, with the development of next generation sequencing (NGS), an enormous amount of omics data has been generated daily to provide a better understanding of the human genome. Through the enormous advances in sequencing technology, we can now identify genetic variants and epigenetic modifications more easily and more cheaply. Population based association studies, such as genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS), provides a reliable and effective way to mine the relationship between diseases and variants based on a statistical test framework. However, these approaches have a few limitations: 1) low coverage for the detection of potential risk variants due to the array design, for example, The Illumina Infinium HumanMethylation450 (450K) array only tests the methylation levels of approximately 480K representative CpGs. This represents a coverage of around 2% CpGs in the whole human genome; 2) in the case of variants with low minor allele frequencies (MAF), association tests are limited in terms of their discovery power. Accordingly, extending the discovery power of association tests by combining current findings from association tests and genome profiles from sequencing becomes an important research focus.

Combining omics data from different data sources (e.g., different databases or projects) and types (e.g., different experiment assays) offers the opportunity to analyze multiple datasets simultaneously for the purpose of discovering novel biological insights, which cannot be accomplished using a single dataset. The main features used in my thesis are obtained and processed from two large public databases at the national level: The NIH Roadmap Epigenomics Mapping Consortium (REMC)[1] and The Encyclopedia of DNA Elements (ENCODE)[2]. These databases contain omics profiles across multiple experiment assays, such as, formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq)[3], DNase I hypersensitive sites sequencing (DNase-seq)[4], Histone Modification Chromatin immunoprecipitation sequencing (ChIP-seq)[5], Transcription Factor (TF) ChIP-seq ChIP-seq[5], PolyA

RNA-seq[6], Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)[7], which facilitate the detection of inherent correlations between various features. Different experiment assays or sequencing techniques provide different insights. For example, DNase-seq and FAIRE-seq can help identify the accessible DNA regions in the genome. RNA-seq quantifies the abundance of RNA to measure the activity of transcription. ATAC-seq helps to assess genome-wide chromatin accessibility. Also, the omics profiles in these databases are derived from various cell types and tissues, adding another dimension to understanding how genome activity affects human health.

Machine learning is a method of learning from the pattern and forming the generalized rule for decision making in future applications. It is more and more widely applied in the area of life science, especially for the functional annotation of genetic and epigenetic variants. Some of them are designed for the prediction of general functionality or general pathogenicity of variants, for example, logistic regression was used in CADD[8] that prioritized functional, deleterious, and pathogenic variants. Random forests were used in GWAVA[9] to distinguish disease-implicated variants from benign variants. Some of them are designed in a more disease specific way, for example, a hybrid two-stage model with support vector machine, random forests, logistic regression, the Lasso[10] and elastic net was used in BioMM[11] to identify epigenetic signatures of schizophrenia. Tree based ensemble model was used in DIVAN[12] to identify specific disease associated single nucleotide polymorphisms (SNPs) for approximately 50 individual diseases. The collection of more extensive features and the development of more powerful machine learning approaches have pushed the potential for more accurate variant annotation to its limit.

1.2 Outline of the dissertation

There are three chapters in my thesis and three machine learning predictive models are proposed for tackling different disease risk annotation problems.

EWASplus is tool that aims to extend the current coverage of The Illumina Infinium HumanMethylation450 (450K) to the whole human genome. It learns from the array-based EWAS results generated using 450K array with multiple base classifiers for ensemble learning and then generates risk scores by applying the prediction model genome-wide. EWASplus has been tested on EWAS studies conducted on Alzheimer's disease (AD), a progressive neurodegenerative disease. The original EWAS results are obtained from various brain tissues (prefrontal cortex, middle temporal gyrus, etc.) and different brain data cohorts (ROS/MAP. London, Arizona, Mount Sinai). Ideally, EWASplus can be applied to any EWASs as long as the original EWAS summary statistics are available such as significance p-values.

CASAVA (Disease Category-specific Annotation of Variants) aims to provide a middle ground between disease-neutral annotation and disease-specific annotation. It pools risk variants from related diseases belong to the same category together to overcome the problem of lack of positive training samples. As a result, CASAVA is able to provides predictions for 24 major disease categories at 200-bp resolution for the entire genome.

DRAFT (Disease Risk Annotation with Few shoTs learning) is an end-to-end deep learning-based approach that incorporates contrastive learning to tackle the lack of risk variants that hinder the application of traditional deep learning models to this research field. In addition, it leverages the recent development of powerful deep learning models and mitigate the problem of lacking sufficient number of high-quality positive training samples.

Chapter 2 *EWASplus*: An ensemble learning approach for the risk prediction of Alzheimer's Disease associated CpGs

2.1 Introduction

Alzheimer's disease (AD) is influenced by both genetic and environmental factors; thus, brain epigenomic alterations may provide insights into AD pathogenesis. Multiple array-based Epigenome-Wide Association Studies (EWASs) have identified robust brain methylation changes in AD; however, array-based assays only test about 2% of all CpG sites in the genome. Here, we develop EWASplus, a computational method that uses a supervised machine learning strategy to extend EWAS coverage to the entire genome. Application to six AD-related traits predicts hundreds of new significant brain CpGs associated with AD, some of which are further validated experimentally. EWASplus also performs well on data collected from independent cohorts and different brain regions. Genes found near top EWASplus loci are enriched for kinases and for genes with evidence for physical interactions with known AD genes. In this work, we show that EWASplus implicates additional epigenetic loci for AD that are not found using array-based AD EWASs.

2.2 Background

Alzheimer's disease (AD) is an age-dependent, neurodegenerative disorder, the leading cause of dementia, and a major public health concern world-wide[13]. AD is a complex illness due to environmental and genetic factors with a heritability of ~70%[14, 15]. Compared to genome-wide association studies (GWASs), there are relatively fewer studies examining AD-associated epigenetic changes in the human brain. Yet, understanding epigenetic changes in the brain is important because they will likely illuminate both heritable and environmental aspects of AD pathogenesis. One of the most well described epigenetic changes, DNA methylation (DNAm), is strongly linked with transcription regulation[16], is heritable[17], and notably changes in response to environmental exposure[18-20], such as smoking[21, 22]. Important for AD and other age-dependent illnesses, it is also known to change with age[23].

Epigenome-wide association studies (EWASs) use array-based assays to test whether DNAm at particular CpG sites (abbreviated CpGs hereafter) is associated with a disease[24-27]. Multiple AD EWASs have identified differential DNAm associated with AD in different regions of the human brain, including prefrontal cortex (PFC)[28], entorhinal cortex (EC), superior temporal gyrus (STG), cerebellum (CER)[29], temporal pole region, temporal cortex, glia, neuron nuclei, non-neuronal nuclei[30], and superior temporal gyrus[31]. These works revealed AD-associated differential DNAm such as those near ANK[28, 31] and CDH23[28, 29], which are distinct from AD GWAS signals. Although these studies have identified new AD-associated genes, array-based methods are limited because they only test about 2–3% of all CpGs in the human genome and have known technical limitations[32]. To overcome these challenges, we tested whether a machine learning approach could be used to identify additional AD-associated CpGs on a genome-wide scale.

In this work, we construct a supervised machine learning (ML) binary classifier named EWASplus to identify CpGs associated with AD. Given that epigenetic features and DNAm status are interconnected, we hypothesize that we can identify AD associated CpGs using genomic and epigenetic features. Training data are derived from array-based EWASs, and features include relevant genomic and epigenomic profiling data (e.g., chromatin accessibility, histone modifications). After model training, we apply the trained model to the entire genome to identify additional AD-associated CpGs. Finally, we perform targeted bisulfite sequencing experiments to validate our in-silico predictions. We find the highest rate of AD association for regions harboring putative CpGs predicted by EWASplus (65.8%; 25 out of 38), follow by CpGs known to associate with AD by methylation arrays (60.0%; 6 out of 10). Experimental validation shows predicted CpGs are 2.2 times more likely to be associated with AD (1.00×10^{-9}) than negative control CpGs. These results suggest EWASplus is capable of providing credible information to identify additional AD-associated CpGs.

2.3 Results

2.3.1 EWASplus overview

The goal of EWASplus is to identify additional disease associated CpGs that are not included on the methylation arrays. Currently, the most popular methylation arrays only represent 2–3% of all CpGs in the human genome. EWASplus aims to increase the number of CpGs tested in EWASs to a genome-wide scale. A comparison of CpG coverage between the 450K methylation array and EWASplus is shown in Figure 2-1. Standard EWAS operates under a testing framework, but EWASplus frames the problem as a supervised learning (i.e., classification) framework. The EWASplus approach (Figure 2-2) is to (1) use summary statistics from array-based EWASs to classify all CpGs on the array into either trait-associated (positive) or neutral (negative) group; (2) perform feature selection to identify the most informative features from a collection of 2256 genomic and epigenomic annotations; (3) train an ensemble learning model capable of identifying CpGs for trait association; and (4) score all CpGs in the entire genome to identify additional trait-associated CpGs not present on the array.

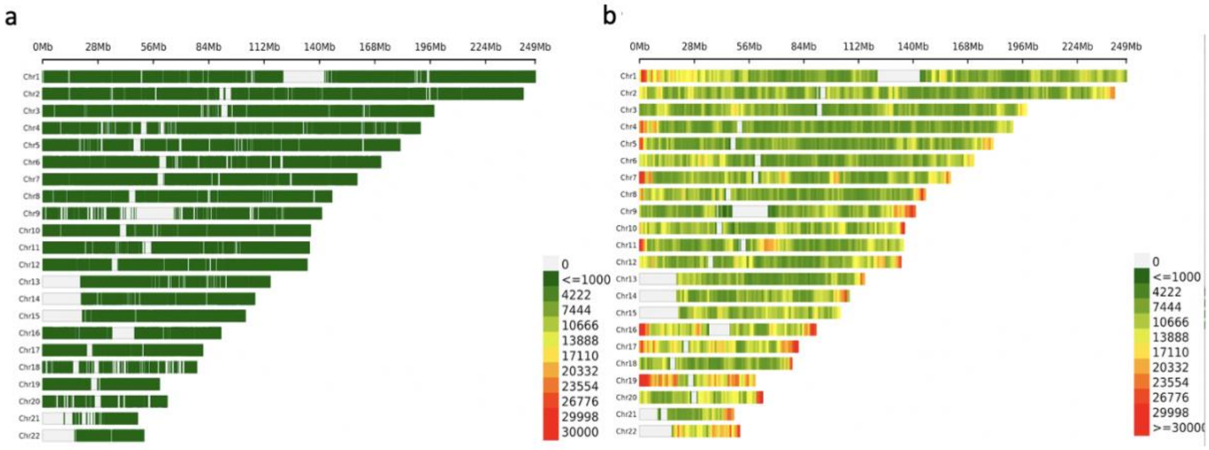


Figure 2-1 Comparison of methylation coverage between Illumina 450K methylation array and EWASplus.

a. The density of CpGs covered by Illumina 450K methylation array. b. The density of CpGs covered by EWASplus. The figure legend for both subplots have the same color bar scale. The darker red indicates a higher CpG density and the darker green means a lower CpG density.

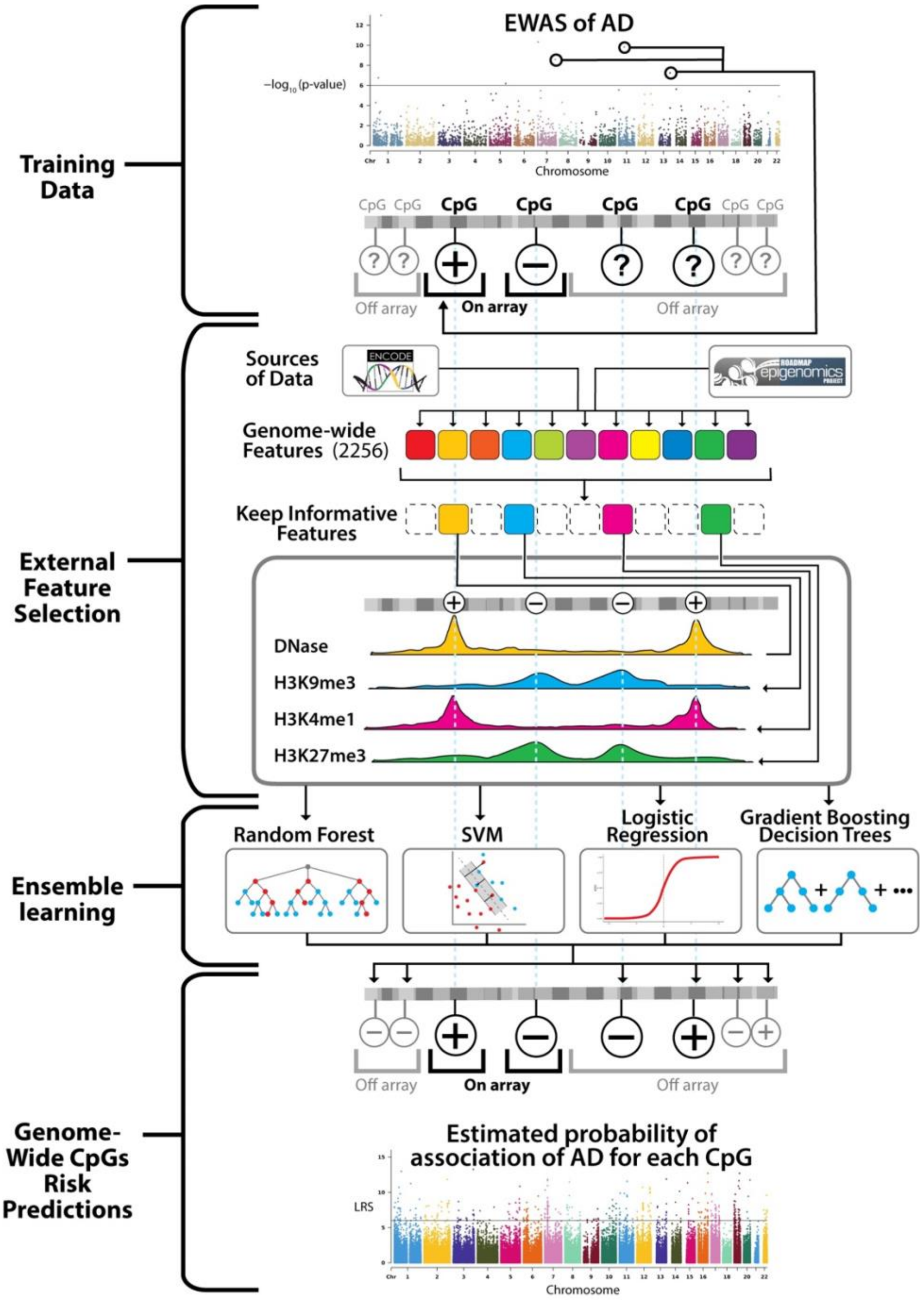

Training Data
External Feature Selection
Ensemble learning
Genome-Wide CpGs Risk Predictions
EWAS of AD
−log10(p-value)
Chromosome
CpG CpG CpG CpG CpG CpG CpG CpG
Off array
On array
Off array
Sources of Data
ENCODE
ROADMAP epigenomics PROJECT
Genome-wide Features (2256)
Keep Informative Features
DNase
H3K9me3
H3K4me1
H3K27me3
Random Forest
SVM
Logistic Regression
Gradient Boosting Decision Trees
Off array
On array
Off array
Estimated probability of association of AD for each CpG
LRS
Chromosome

Figure 2-2 Overview of EWASplus approach.

The EWASplus procedure is composed of four major steps: (1) Training data collection from existing EWASs; (2) External feature (from sources such as ENCODE and Roadmap Epigenome consortia) selection; (3) Ensemble learning; and (4) Genome-wide CpGs risk prediction, in which trained ensemble learning model is applied genome-wide to score all CpGs.

To prepare the training set, EWASplus gathers the most significant CpGs identified from array-based EWAS to form a positive training set. To reflect the fact that there are far fewer significant trait-associated CpGs in the genome than the trait-neutral ones, EWASplus selects a matching negative training set with similar genomic context that is ten times larger than the positive training set.

EWASplus employs an ensemble learning strategy and four different methods were chosen as the base learner: regularized logistic regression (RLR), support vector machine (SVM) classifier, random forest (RF), and gradient boosting decision trees (GBDT). To identify the best ensemble model, we tested all possible combinations of these base learners and found that the combination of RLR and GBDT gives the best performance overall, and hence was selected to be the ensemble model in this study. RLR has the best recall but relatively low precision, while GBDT has the best precision but relatively low recall. When these two models are ensembled, the underfitting property of RLR can effectively offset the overfitting from GBDT while keeping enough model complexity. More detailed description can be found in "2.4.7 Performance evaluation metrics" in the Methods section.

EWASplus can be applied to any array-based EWAS to extend its coverage. In this study, we tested EWASplus on data collected from four different cohorts: ROS/MAP (sample size 717), London (sample size 113), Mount Sinai (sample size 146), and Arizona (sample size 302). All original EWASs were performed using the Illumina 450K methylation array.

2.3.2 EWASplus performance compared to methylation array

To evaluate the performance of EWASplus, we first considered its performance on CpGs present on the Illumina 450K methylation array (henceforth referred to as the "array"). Given the large sample size (n = 717), we choose data from the ROS/MAP cohort as the main dataset for performance evaluation. Methylation is measured on DNA derived from post-mortem PFC. Standard EWAS were conducted on six different AD-related traits: beta-amyloid density, Braak staging, the Consortium to Establish a

Registry for Alzheimer's Disease (CERAD) score, cognitive trajectory, global AD pathology, and neurofibrillary tangle density. We trained a separate classifier for each of the six traits.

EWASplus results are summarized in Table 2-1 (see section 2.4.6 Hyperparameter tuning and ensemble model for detailed description of the approach of evaluation). The area under the receiver operator characteristic (ROC) curve (AUC) values from the six traits range from 0.831 (cognitive trajectory) to 0.962 (neurofibrillary tangles) (Figure 2-3a). The area under the precision-recall curve (PRC) (AUPRC) values from the six traits range from 0.502 (CERAD) to 0.858 (neurofibrillary tangles) (Figure 2-3b). These results indicate that EWASplus works well to predict significant AD-associated CpGs for methylation measured by the array. Among the six traits, we observe the best performance for neurofibrillary tangles.

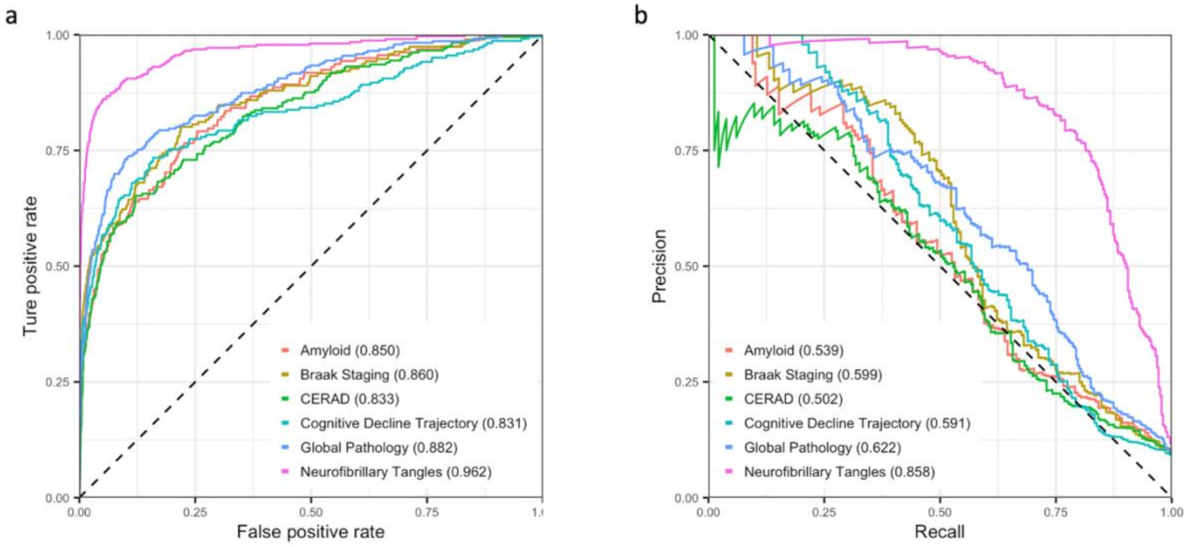


Figure 2-3 Summary EWASplus results.

a ROC curves of the predictive performance of EWASplus on the six traits in the ROS/MAP cohort. b Precision-recall curves of the predictive performance of EWASplus on the six traits in the ROS/MAP cohort.

Outcome	Outcome Type	AUC	AUPR	F1	Precision	Recall
Beta-amyloid	Pathologic, IHC	0.850	0.539	0.492	0.423	0.589
Braak Staging	Pathologic, Silver Stain	0.860	0.599	0.530	0.487	0.581
CERAD	Pathologic, Silver Stain	0.833	0.502	0.508	0.457	0.571
Cognitive Trajectory	Clinical	0.831	0.591	0.516	0.451	0.604
Global Pathology	Pathologic, Silver Stain	0.882	0.622	0.577	0.507	0.671
Neurofibrillary Tangles	Pathologic, IHC	0.962	0.858	0.754	0.677	0.852

Table 2-1 Summary of performance evaluation of all six AD related traits.

To further evaluate EWASplus, we asked whether the EWASplus prediction score is capable of distinguishing CpGs with differential DNAm between AD case and control status. To answer this question, we selected four groups of CpGs that differ with respect to differential DNAm association with AD: (a) AD-associated CpGs in the positive training set (i.e., p-value less than the EWAS threshold); (b) CpGs suggestively associated with AD (i.e., p-value slightly greater than the EWAS threshold); (c) CpG not associated with AD but not in negative training set; and (d) CpGs not associated with AD and in the negative CpG training set (i.e., p-value greater than the EWAS threshold and in negative training set). On average, we find a significant difference for EWASplus prediction scores between suggestively positive and negative CpGs that are not in the training sets (Wilcoxon rank-sum test; $p < 3.64 \times 10^{-16}$) for all six traits. Scores for CpGs in group b are similar to those in group a (positive training set), albeit with higher variation, whereas scores in group c have almost the same scores as group d (negative training set). As expected, our results demonstrated excellent capability of EWASplus in distinguishing CpGs that show AD association or not.

2.3.3 EWASPlus performance for off-array CpGs

We applied the six classifiers trained on the six AD traits using EWASplus to the entire human genome to obtain a prediction score for every CpG (Figure 2-4a). The top ten CpGs with the highest composite scores are listed in Table 2-2. The total number of CpGs with a prediction score is about 78 times the number of CpGs present on the Illumina 450K methylation array. The prediction scores for all CpGs are provided at the EWASplus Github site.

Chr	Position (bp)	Beta-Amyloid	Braak Staging	CERAD	Cognitive Decline	Global Pathology	Neurofibrillary Tangles	Genes within 50 kb of associated CpG
7	**27148225**	**5.605**	**5.037**	**3.799**	**4.495**	**4.612**	**7.424**	***HOTAIRM1, HOXA-AS2, HOXA-AS3, HOXA1, HOXA2, HOXA3, HOXA4, HOXA5, HOXA6, HOXA7***
5	172175606	4.679	4.662	5.009	4.658	6.248	3.720	*DUSP1*
7	47367933	4.008	4.736	5.251	4.254	5.278	5.210	*TNS3*
19	**46270392**	**4.462**	**4.447**	**6.311**	**4.362**	**4.937**	**3.727**	***FBXO46, SIX5, DMPK, DMWD, RSPH6A, SYMPK***
6	35286078	5.141	5.947	4.497	3.731	3.762	5.072	*ZNF76, DEF6, PPARD*
19	10736075	4.130	5.977	3.387	4.047	4.400	5.845	*AP1M2, SLC44A2, ILF3, ILF3-AS1*
9	116225986	3.408	4.977	4.253	3.630	5.307	5.893	*C9orf43, RGS3*
1	59280358	3.065	6.248	3.973	5.179	5.130	3.755	*LINC01135, JUN*
19	**15563592**	**6.579**	**4.549**	**3.132**	**3.849**	**4.009**	**5.215**	***MIR1470, AKAP8L, WIZ, RASAL3, PGLYRP2***
7	151433271	3.577	4.432	4.265	4.739	4.743	4.814	*PRKAG2*

Table 2-2 Top ten CpG loci for six AD-relevant traits. Genes within 50 Kbp of the region are provided. Genes with prior evidence of being associated with AD given in bold.

For the off-array CpGs, we examined the distribution of prediction scores for different types of genomic regions. We hypothesized that top CpGs with the highest prediction scores would be located in functional regions such as enhancers and promoters, and we find this to be the case (Figure 2-4b). The normalized proportion for enhancers ranges from 15.73 to 43.19% and exons range from 4.69 to 23.99%, which are both significantly higher than the expected occurrence of these regions in the high prediction score percentile intervals (binomial test for the highest prediction score quantile interval: $p < 1.00 \times 10^{-99}$ for both enhancers and exons). To better understand the properties and context of top-ranked CpGs predicted by EWASplus, we selected the top 10k CpGs with the highest overall EWASplus prediction scores and analyzed their chromatin states (15-state model) defined in dorsolateral prefrontal cortex. We calculated the enrichment (or depletion) of the 15 chromatin states in the top 10k CpGs. As a result, we found that all six AD-related traits are enriched for sites annotated as flanking active transcription start site (TSS) (binomial test; $p < 1.00 \times 10^{-99}$ for all traits), active TSS (binomial test; $p < 1.00 \times 10^{-99}$ for all traits), enhancers (binomial test; $p < 1.22 \times 10^{-9}$ for all traits), and repressed PolyComb (binomial test; $p < 1.00 \times 10^{-99}$ for all traits), and under-represented for sites within quiescent regions (binomial test; $p < 1.00 \times 10^{-99}$ for all traits) (Figure 2-4c). There is no significant difference in the enrichment patterns across the six AD traits. These results support the conclusion that top CpGs associated with AD tend to be located in functional regions.

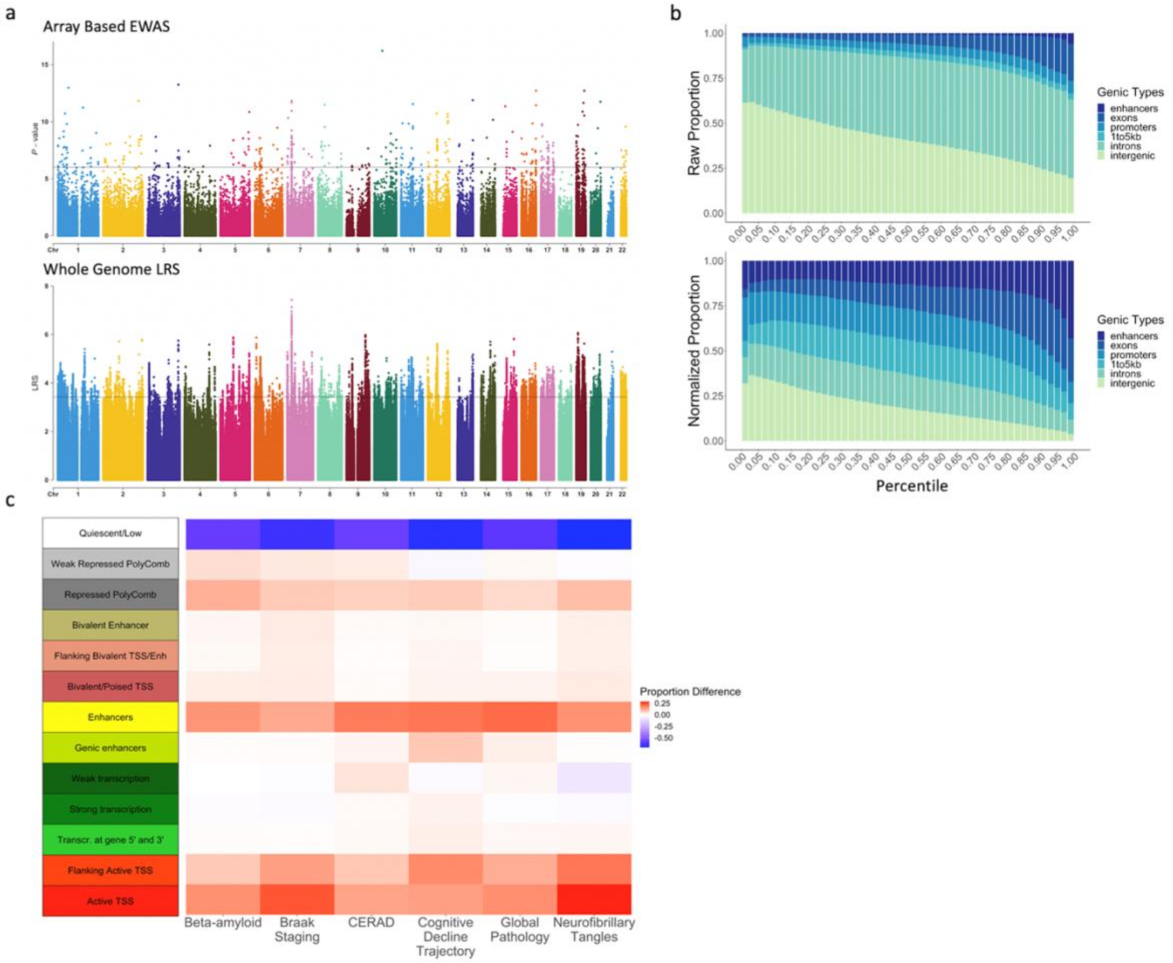


Figure 2-4 Genome-wide prediction results.

a Manhattan plots for neurofibrillary tangles: the top panel is for on-450K CpGs with EWAS p-values and the bottom panel is for whole-genome CpGs with imputed LRS by EWASplus. The y-axis is the log-scale rank scores. The top-ranked CpG has the LRS of 7.42 (about empirical p-value of 3.8×10^{-8}); the top 100th ranked CpG has the LRS of 5.42 (about empirical p-value of 3.8×10^{-6}) and the top 10,000th ranked CpG (about empirical p-value of 3.8×10^{-4}) has the LRS of 3.42. b Raw and normalized stacked-proportion histograms for different genomic annotation types. Source data are provided as a Source data file. c The difference of observed and expected chromatin states proportion for the top 10,000 loci across the six AD-related traits: Beta-amyloid, Braak staging, CERAD, cognitive trajectory, global pathology, and neurofibrillary tangles. Source data are provided as a Source data file. The annotated chromatin states are from Roadmap Epigenetics Project, and we used the core 15-state model chromatin states for the dorsolateral prefrontal cortex tissue type. To minimize ambiguity, we require only a single annotation type is assigned for each CpG site. if a CpG has multiple annotations, we only record the most "significant" annotation with the following order: enhancer > promoter > exon > intron > near gene (1–5 kb to the TSS) > intergenic. We do not list 5′ UTR and 3′ UTR since these two types are within the first and last exon of each gene according to the UCSC annotation system.

2.3.4 Comparison with a competing method

In a recent work, using array-measured methylation levels, Zhang et al.[33] develop a computational algorithm to impute the methylation levels on CpG sites genome-wide including those not on the Illumina 450K array. Their approach employed about 125 genomic and epigenomic features (the number varies when including different sets of individual-level features) mainly composed of regulatory marks from ENCODE project. Although not designed for trait association prediction, one could apply this method to impute methylation levels for every individual sample and on every CpG site. Subsequently, association test can be conducted on these imputed methylation measures to identify CpGs significantly associated with a trait of interest.

To compare such a strategy with EWASplus, we applied Zhang et al.'s method and used the imputed methylation values to conduct an association test. We found that the AUC for EWASplus is between 0.178 and 0.329 higher compared to the adapted Zhang et al. approach; AUPR for EWASplus is 0.219 to 0.364 higher than adapted Zhang et al. approach across six AD- related traits (Figure 2-5).

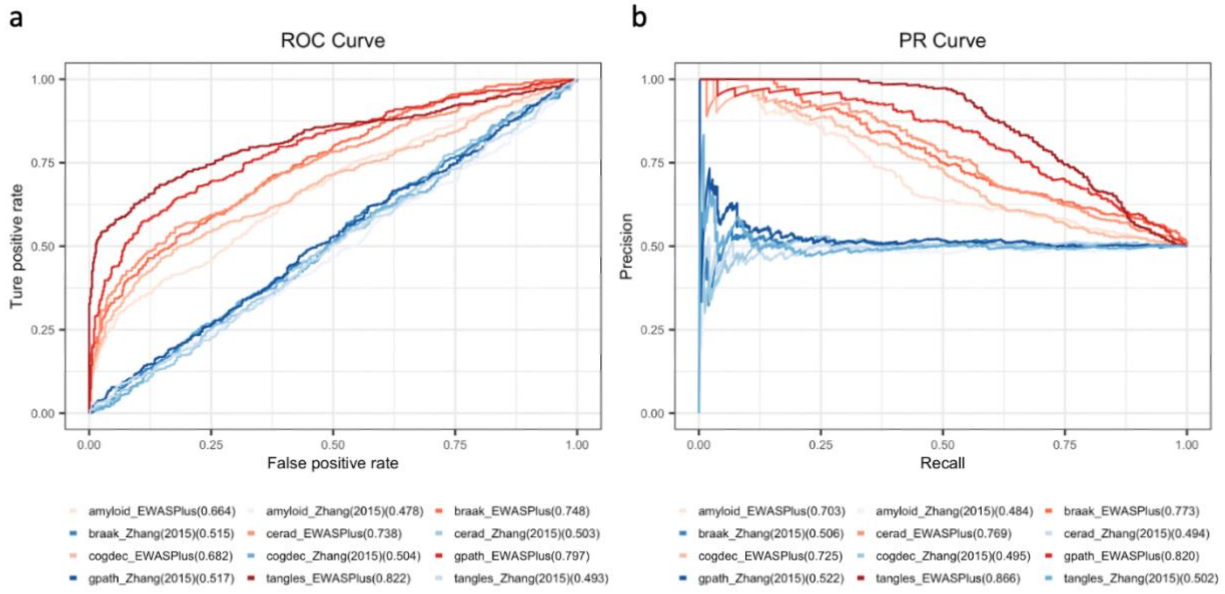


Figure 2-5 Performance comparison between EWASplus and adapted Zhang et al. methylation level imputation method.

a. Receiver Operating Characteristic curves of the predictive performance of EWASplus versus adapted Zhang et al. method. b. Precision-Recall curves of the predictive performance of EWASplus versus adapted Zhang et al. method. Source data are provided as a Source Data file.

2.3.5 Experimental validation of EWASplus predictions

To experimentally test the validity of the prediction scores reported by EWASplus, we performed targeted bisulfite sequencing to measure the methylation level at 559 selected CpGs from 150 randomly selected participants from the Religious Orders Study (ROS) or Memory and Aging Project (MAP) cohorts who are representative of both studies and have available brain tissue for bisulfite sequencing (Table 2-3). CpGs were selected for independent validation from the top EWASplus predicted sites using a stepwise selection process that prioritized regions with the highest predicted scores that were physically separated by at least 500 bp. For comparison purposes, we also randomly selected CpGs from regions with predicted scores in the lower half but similar physical characteristics (e.g., GC content). In addition, we targeted CpGs on the array that could serve as positive controls. After quality control, 319 CpGs were analyzed including 31 CpGs on the 450K array identified as AD-associated[34], 260 off-array CpGs predicted to be AD-associated based on EWASplus, and 28 off-array CpGs predicted to not be AD-associated. These 319 CpGs can be grouped into 58 independent clusters (referred to as CpG cluster hereafter) on the genome that belongs to three groups: 38 off-array predicted AD-associated, 10 on-array AD-associated, and 10 off-array predicted not AD-associated. For performance comparison, we combined test results from the six individual traits. Due to the limited sample size, we call a CpG cluster AD-associated if at least one of the CpGs at the locus achieves unadjusted p-value for differential DNAm < 0.05 for any of the six traits. Similar to our results from individual traits, we found that positive CpG clusters predicted by EWASplus have the highest rate of association with at least one AD trait (65.8%, or 25 of 38), followed by CpG clusters identified by array-based EWAS (60.0%, or 6 of 10). In contrast, the negative control CpG clusters predicted by EWASplus have the lowest (30.0%, 3 of 10) (Table 2-4). Thus, CpGs with top EWASplus scores are about 2.2 times more likely to be associated with an AD trait (Binomial test, $p < 1.00 \times 10^{-9}$).

	Sequenced (N=150)	**Un-sequenced (N=589)**
Male	46	223
Female	104	366
Age of death	87.569 (6.408)	88.096 (6.738)
Education	16.080 (3.639)	16.487 (3.566)
Amyloid	3.372 (3.992)	3.498 (3.619)
Braak Staging	3.433 (1.297)	3.414 (1.258)
CERAD	2.407 (1.193)	2.299 (1.148)
Cognitive Decline Trajectory	-0.026 (0.107)	-0.030 (0.107)
Global Pathology	0.693 (0.660)	0.705 (0.615)
Neurofibrillary Tangles	7.481 (9.637)	6.102 (7.599)

Table 2-3 Demographic information of the sequenced samples and un-sequenced samples from original EWAS ROS/MAP cohort.

	# of positives in EWASplus predicted positives (%) total = 38	# of positive in on-array positives (%) total = 10	# of positives in EWASplus predicted negatives (%) total = 10
Any Trait	25 (65.8)	6 (60.0)	3 (10.0)
Beta-Amyloid	17 (44.7)	1 (10.0)	2 (20.0)
Braak Staging	11 (28.9)	2 (20.0)	1 (10.0)
CERAD	17 (44.7)	2 (20.0)	3 (30.0)
Cognitive Trajectory	7 (18.4)	3 (30.0)	0 (0.0)
Global Pathology	13 (34.2)	2 (20.0)	3 (30.0)
Neurofibrillary Tangles	16 (42.1)	5 (50.0)	1 (10.0)

Table 2-4 Comparison of number and proportion of differentially methylated CpGs in various categories of CpGs. Methylation level is measured by Targeted Bisulfite Sequencing Experiment.

2.3.6 EWASplus performance on multiple cohorts

To further test EWASplus, we examined its performance using data from three additional cohorts: London cohort[29] (prefrontal cortex, N = 113), Mount Sinai cohort[35] (prefrontal cortex, N = 146), and Arizona cohort[36] (middle temporal gyrus, N = 302). In all three studies, Braak staging (treated as a continuous variable) is used as the trait in the EWAS studies, as described in Smith et al.[35]. Detailed information about these cohorts is summarized in Table 2-5.

	Braak Stage	**N**	**Gender (M/F)**	**Age of death**
London (N=113)	0-II	29	13/16	77.6(12.8)
	III-IV	18	7/11	88.5(5.2)
	V-VI	66	26/40	85.4(8.1)
Mount Sinai (N=146)	0-II	60	32/28	82(7.6)
	III-IV	42	12/30	88.8(6.6)
	V-VI	44	12/32	88.0(7.5)
Arizona (N=302)	0-II	61	40/21	80.3(8.2)
	III-IV	97	50/47	86.9(6.9)
	V-VI	144	63/81	82.3(8.5)
ROS/MAP (N=739)	0-II	151	75/76	83.6(7.2)
	III-IV	423	148/275	88.8(6.3)
	V-VI	165	46/119	89.8(5.2)

Table 2-5 Cohort Characteristics.

We found that EWASplus performed well in all three datasets. The AUC values range from 0.697 (London 1) to 0.863 (Mount Sinai). The AUPRC values range from 0.233 (Arizona) to 0.604 (Mount Sinai). The complete results including all evaluation metrics can be found in Table 2-6.

Cohort	Brain Tissue	AUC	AUPR	F1	Precision	Recall
London	Prefrontal Cortex	0.697	0.272	0.325	0.248	0.471
Mount Sinai	Prefrontal Cortex	0.863	0.604	0.481	0.364	0.708
Arizona	Middle temporal gyrus	0.699	0.233	0.275	0.196	0.461

Table 2-6 Summary of performance evaluation on three additional cohorts of samples: London, Mount Sinai and Arizona.

To understand the most relevant factors influencing EWASplus performance among the different datasets, we treated the performance measurement testing AUC as the response variable and tested numerous independent variables using the linear regression model. We found that when choosing the positive EWAS threshold (negative logarithm transformed p-values) as the independent variable, simple linear regression achieved R^2 of 0.588 using other performance measures such as AUPRC and F1 values produced similar results. These results suggest that perhaps the most relevant factor that influences EWASplus performance is the quality and power of the original EWAS, which depends on the effect and sample sizes.

2.3.7 Biological insights into AD

To glean biological insights from the EWASplus results, we examined genes surrounding some of the highest EWASplus scoring CpGs. Interestingly, we found that the highest scoring CpG is located inside the HOXA gene cluster, which has been identified by three independent array-based EWASs of cortical brain tissue associated with Braak staging, a measure of neurofibrillary tangles[29, 35, 37]. In contrast to prior analyses that identified individual HOX genes, EWASplus results identify a 40 kb region on chromosome 7 that includes multiple homeobox genes, e.g., HOXA2, HOXA3, HOXA4, HOXA5, and HOXA6, that are associated with AD (Figure 2-6).

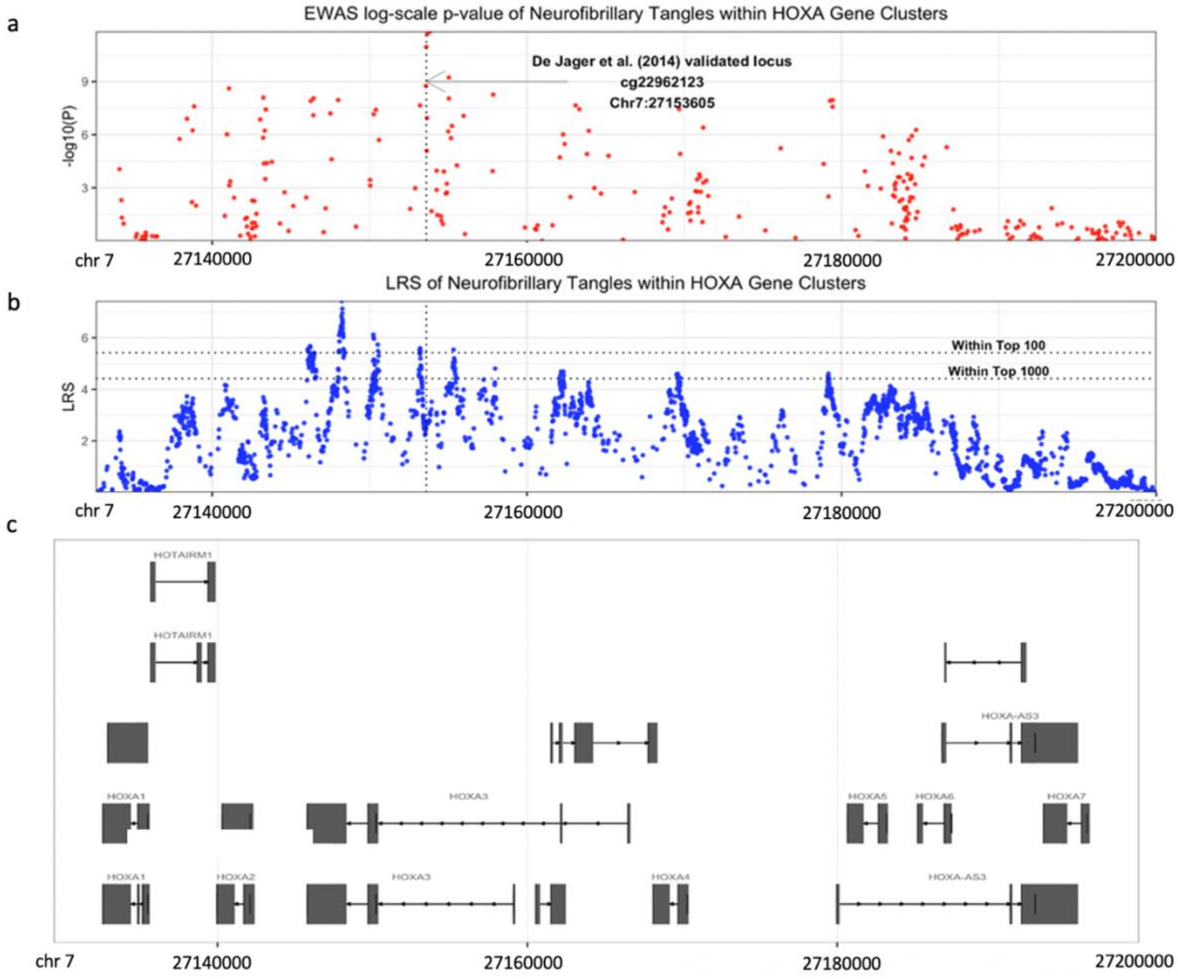


Figure 2-6 Manhattan plot of neurofibrillary tangles EWAS at the HoxA locus on chromosome 7.

a. Array-based EWAS p-values. The most significant CpG identified by De Jager et al. are shown with an arrow. b. EWASplus predicted LRS. c. The landscape of the HoxA cluster genes.

In addition, of the top 10 detached EWASplus scoring CpGs, seven were not previously implicated in any EWAS of AD. Here detached means any two CpGs on this list are at least 10 kb away from each other. Gene set enrichment analysis by GeNets[38] using all genes located within 5 kb of the top 100 EWASplus scoring CpGs (123 genes) revealed a significant enrichment of protein kinases ($p = 0.044$) — ALPK3, DMPK, MAP3K11, MAP4K1, and TAOK328. Identification of kinases within AD is of particular interest given that neurofibrillary tangles, a hallmark neuropathology of AD, result from hyperphosphorylation of microtubule-associated protein tau (MAPT)[39]. In addition, we found that genes within the top EWASplus regions have evidence of physical interaction with known AD genes or AD GWAS loci (e.g., PRKAG2 and TNS3 interact with APOE, CLU, APP, PSEN1/2, and RIN2 and RIN3 interact with BIN1). These analyses support the idea that EWASplus is able to identify interesting underlying biological relationships in AD.

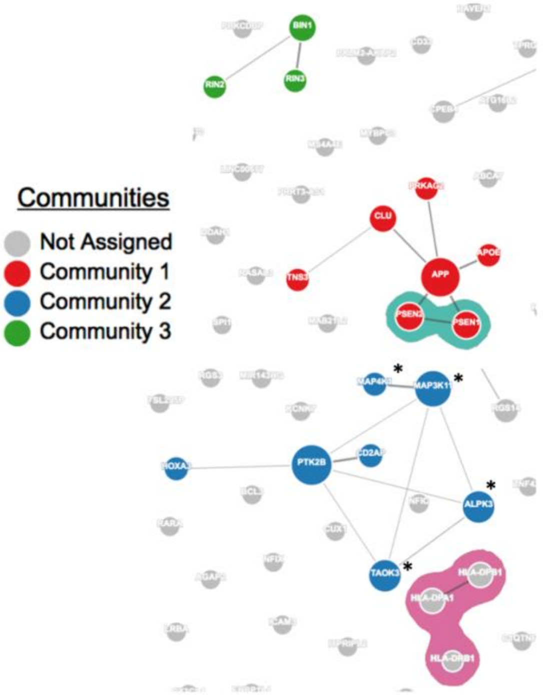


Figure 2-7 Selected protein-protein interaction (PPI) networks and communities among known AD GWAS genes (n=28) and top AD EWASplus genes (n=123).

The lines represent physical PPIs between proteins. The thickness of the lines is proportional to the evidence for the PPI. The black asterisk indicates genes that are known kinases.

2.4 Methods

2.4.1 Cohorts

The main dataset used in this study comes from the ROS/MAP cohorts. ROS and MAP are longitudinal cohort studies of aging and AD led by investigators at the Rush Alzheimer's Disease Center[40, 41]. Participants give written informed consent for annual assessments, signed an Anatomic Gift Act, and a repository consent to allow their data and biospecimens to be repurposed. Each year, participants undergo a detailed medical, neurological, and neuropsychiatric assessment. After death, each participant undergoes a detailed brain autopsy with neuropathologic examination. Both ROS and MAP were approved by the Institutional Review Board of Rush University Medical Center. They share a large common core of data at the item level to allow efficient merging of datasets. ROS/MAP resources can be requested at https://www.radc.rush.edu.

In addition, we also obtained data from three separate cohorts: London, Mount Sinai, and Arizona. The "London" cohort refers to prefrontal cortex tissue obtained from 113 individuals archived in the MRC London Neurodegenerative Disease Brain Bank. The details of the cohort are described in Lunnon et al.[29]. The "Mount Sinai" cohort refers to prefrontal cortex tissue obtained from 146 individuals archived in the Mount Sinai Alzheimer's Disease and Schizophrenia Brain Bank. Details of this cohort is described in Smith and colleagues[42]. The "Arizona" cohort refers to 302 middle temporal gyrus samples from The Sun Health Research Institute Brain Donation Program[36]. The details of this cohort are described in Brokaw et al.[43].

2.4.2 Sample preparation and differential DNAm CpGs identification

DNAm data were generated from dorsolateral PFC (Broadman area[44]) of post-mortem samples obtained from individuals in the ROS/MAP cohorts.

DNAm profiling was performed with the Illumina HumanMethylation450 Beadchip array[28]. After excluding non-Caucasian subjects, 717 ROS/MAP participants with array DNAm data remained for analysis. We obtained raw IDAT files from the Synapse website (Synapse ID: syn7357283) and removed probes annotated to multiple chromosomes or the X and Y chromosomes by Illumina, probes that cross-hybridize with other probes due to sequence similarity (identified by Chen et al.[45]), probes with a detection $p > 0.01$ in any sample, probes without a CpG, and probes that overlap with a common SNP (identified by Barfield et al.[46]). After this filtering, a total of 334,465 autosomal CpGs remained for analysis.

For the EWAS analyses, each probe was normalized using the BMIQ algorithm from the Watermelon R package, and adjusted for batch effects using the ComBat function from the sva R package[47]. We used the CpGassoc[48] R package to test if the methylation level of each array CpG is associated with the trait of interest via regression methods. All models were adjusted for proportion of neurons, age at death, sex, post-mortem interval, plate, study, and years of education. Neurons were added as a covariate to avoid potential confounding due to differences in the cellular composition of the tissue samples. The proportion of neurons in each sample was estimated using the CETS R package and reference methylation data from isolated neuronal nuclei[49].

We performed EWASs for the following six AD-related traits: (1) beta-amyloid load which is the percent area of beta-amyloid based on image analysis; (2) neurofibrillary tangle density by stereology; (3) CERAD score; (4) Braak stage; (5) global AD pathology burden; (6) cognitive trajectory based on the average z-score of 17 cognitive function tests. Beta-amyloid and neurofibrillary tangle were measured in the cortex using immunohistochemistry with antibodies specific to beta-amyloid and phosphorylated-tau, as described[41]. We used square-root- transformed values for both traits to improve their normality. CERAD score and Braak stage are semi-quantitative measures that reflect both a neuropathologist’s opinion of AD diagnosis and the distribution and amount of silver-stain-identified neuritic and diffuse plaque and neurofibrillary tangle pathologies, respectively[50-52]. CERAD scores can take on values from one to four indicating definite AD, probable AD, possible AD, and no AD, respectively. CERAD was treated as a continuous trait. Braak stages can take on values from one to six, indicating the increasing spread of neurofibrillary tangle pathology in the brain, and Braak was coded as a binary trait with stages one to three as controls and stages four to six as affected. Global AD pathology burden is a summary measure of silver-stain- identified neuritic plaque, diffuse plaque, and neurofibrillary tangle pathologies[41]. As global AD pathology burden has a skewed distribution, we used square-root-transformed values. Cognitive trajectory, or the rate of change in cognition over time, was estimated for each ROS/MAP participant using a linear mixed model[34]. For each person, cognitive trajectory was estimated as the person-specific random slope of a linear mixed model that included global cognitive function as the longitudinal outcome[53], follow-up year as the independent variable, and sex, age at enrollment, and years of education as covariates.

For the London, Mount Sinai, and Arizona cohorts, we directly used the processed EWAS results reported in Smith et al.[42]. Details of the sample preparation and differential DNAm CpGs identification have been described in previous studies[29, 36, 43].

2.4.3 Training sets selection

For each trait, positive CpGs in the training set were selected based on association test p-values (threshold ranges from 1.00×10^{-7} to 1.00×10^{-5}). For each positive CpG, ten matching negative CpGs were selected from the Illumina 450K array such that they have similar β-values as the positive CpG, but none is considered significant in any of the EWASs conducted on the six traits. We used a conservative threshold ($p > 0.40$) for being not-significant, and β-values were calculated as the mean values of methylation intensity over 717 ROS/ MAP samples for each CpG on the Illumina 450K array.

2.4.4 Base classifiers

We used four different methods as base classifiers with varying model complexity. The goal of this approach was to select the model with the least error to achieve an optimal overall performance. We used four models that included: (1) RLR with L2-penalty, which alleviates overfitting and feature collinearity; (2) SVM classifier[54], which performs well with linearly non-separable classification, a common feature for real-world problems; (3) RF[55], which is a bagging method with decision tree as base learner; (4) GBDT[56], which differs from RF in that a new tree is added to model to gradually optimize the objective function that was set as log loss. EWASplus uses an accurate and efficient implementation of GBDT from package XGBoost[57].

2.4.5 Feature selection

We assembled a comprehensive collection of 2,256 genomic/epigenomic profiles as well as multiple functional annotation scores as features to be used in the model. Omics profiles include TF and histone ChIP-seq, open chromatin, total RNA-seq, and WGBS. Functional annotation scores include CADD[8], GenoCanyon[58], and Eigen/EigenPC[59].

The moderate size of the training sets (between 1,706 and 3,181) may result in overfitting if all features are included in the training. Thus, we used a dimension reduction/feature selection step before the model training. For each trait, we performed feature selection for each of the four base classifiers: RLR, SVM, RF, and GBDT, respectively. For each base classifier, we selected the top 100 most informative features using the training data. In RLR and SVM, features were ranked based on the weights of the fitted model. For RF, features were ranked based on the Gini impurity measure. For GBDT, features were ranked by the gain metric when fitting the model, or, in other words, the improvement in accuracy brought by a feature to the branches it is on.

Next, we ranked the features by the number of times that this feature was selected by the four base learners as informative. We selected the top 60 features (testing on the number of top features ranges 30–100, 60 was selected because it gave the best performance overall). Features were ranked by the number of methods that select the feature as informative. To break a tie, we introduced a secondary sorting method. For each feature, we conducted the Wilcoxon rank-sum test comparing feature values between positive and negative CpGs, and features were ranked from the most significant to the least significant.

2.4.6 Hyperparameter tuning and ensemble model

We used the Tree-of-Parzan Estimators (TPE) implemented in Hyperopt[60] to adaptively search the hyper-parameter space of each component model (base learner) for the best hyperparameter settings. This model-based hyperparameter tuning method is thought to achieve better performance than random search in terms of both accuracy and efficiency.

The hyperparameter tuning for each component model is conducted separately. In the training dataset, we uniformly up-sample the positive CpGs to match the number of negative CpGs to alleviate the imbalance problem. In the outer CV, the whole dataset of positive and negative CpGs were split into training and testing sets in a nine-to-one ratio in each round. Within each round, the nine folds were further split into threefold to conduct the inner 3-fold CV for hyperparameter searching. The best set of hyperparameters was decided by the highest F1 score and it was then used for the remaining onefold in the outer 10-fold CV. Each one of the ten folds in the outer CV layer is used once as the testing set in a round-robin way so that out-sample predictions cover the whole dataset. We evaluate our model with the out-of-bag estimates for testing error and report the evaluation results in Table 2-1 for the ROS/MAP cohort and Table 2-6 for other additional cohorts.

After the outer 10-fold CV, we then built the ensemble model by selecting the best combination of component models. The out-sample predictions of each base learner from the outer 10-fold CV were aggregated in a soft-voting manner to give the ensemble prediction probabilities in different combinations of component models. Due to the problem of class imbalance, we evaluated the performance of the ensemble models using AUC, AUPR, precision, recall, and F1 score.

2.4.7 Performance evaluation metrics

To assess the performance of EWASplus, we used three classes of evaluation metrics: precision and accuracy, AUC and AUPRC, as well as F1 score. Precision measures the true positive rate of a classifier. Accuracy measures the percentage that a classifier correctly labels test samples. For imbalanced datasets where positive samples are of more interest, precision is preferred over accuracy. PRC is preferred over ROC. The F1 score is another widely used performance measure for imbalanced datasets. It takes into consideration both accuracy and precision by assigning each an equal weight in the following calculation formula: $F_1 = 2 \times (\frac{recall*precision}{recall+precision})$. The focus of F1 score is on the positive samples which is usually under-represented.

2.4.8 Binomial test for enrichment of protein kinases

We selected the top 100 CpGs with the highest EWASplus prediction scores across six AD-related traits in a stepwise forward manner such that any two CpGs in the top 100 list are at least 10 kb away from each other. Next, we searched through the 5 kb neighborhood of these 100 CpGs to retrieve all genes that overlapped, for a total of 123 genes. Among these genes, five are known protein kinases. Given a complete list of human kinases (492 from the autosomes) from Kinase.com (http://kinase.com/human/kinome) and a complete list of human genes (31,684 from the autosomes) from Ensembl (http://grch37.ensembl.org), we conducted an enrichment test using binomial distribution which returned an enrichment p-value of 0.044.

2.4.9 Log-scale rank score (LRS) for prioritizing AD-associated loci

In order to better present the whole-genome prediction result, we sorted the prediction scores of each trait and calculated the log-scale rank score (LRS) for each CpG ($LRS = -\log_{10} \frac{rank}{total\ num\ CpGs}; total\ num\ CpGs = 26{,}573{,}858$). The LRS is similar to a log-transformed empirical p-value. A higher LRS means the CpG is more likely to be associated with the trait.

2.4.10 Loci selection for targeted bisulfite sequencing

Targeted bisulfite sequencing was conducted on selected CpGs (with neighboring CpGs profiled unintentionally, as well) for 150 randomly selected samples from the ROS/MAP cohort. Since most features used in model training having only one value in every 200 bp bin, CpGs within a 200 bp bin tend to have similar prediction scores. In order to select a more representative (less clustered) set of loci for experimental validation, we required any pair of selected CpGs must be at least 500 bp apart. The forward selection process is performed in the stepwise manner, starting from the CpG with the highest total LRS score. Due to the limitation of sequencing primer design, not all loci on the candidate list were selected for bisulfite sequencing. The selection process was stopped when a pre-determined sequencing capacity is reached. For comparison, we selected 38 off-array CpG clusters with high prediction scores, 10 clusters of on-array CpGs listed in de Jager et al.[28] and 10 clusters of off-target negative control CpGs.

2.4.11 Adaption of Zhang et al. for comparison with EWASplus

For the purpose of fair comparison, we selected 1000 CpGs that are not from the training set used by EWASplus. Instead, we selected 500 “near positive” CpGs with p-values just above the threshold and 500 negative CpGs with p-values > 0.40 but not in the negative training set used by EWASplus. Comparison is performed in two steps: (1) predict methylation levels for the 1000 CpGs across the 717 samples used to train EWASplus following instruction in Zhang et al., and (2) perform association test with R package CpGassoc[48] to test for differential methylated based on the predicted methylation level from the first step.

Association testing was performed using the same approach as the array-based methylation with CpGassoc58 with modifications. The methylation levels were modeled as logit transformation of β values ($\log(\beta/1-\beta)$) to stabilize the variance[61]. Next, we grouped adjacent CpGs into clusters and conducted the test for differential DNAm. Due to the limited sample size of the study, we call a CpG cluster differential DNAm if the lowest p-value from the Differential DNAm test is less than an unadjusted p-value 0.05 among all CpGs in the CpG cluster. We adjusted for the following covariates: age of death, sex, years of formal education, post-mortem interval, study, and cell type proportion, which was estimated using the CETS R package[49].

2.4.12 Targeted bisulfite sequencing

Multiplex primers were designed to amplify the identified regions using MPD software[62]. The 200–500 ng purified genomic DNA was used for bisulfite conversion (EpiTect Bisulfite Kit (Qiagen)). The treated DNA were used for PCR amplification and PCR amplicons were further purified and pooled together in equal molar. Mixed amplicons were then purified for libraries preparation and deep sequencing (100× or above) using a MiSeq following standard procedures recommended by Illumina. Image analysis and base calling were performed using standard Illumina pipelines. Quality control was performed in the same fashion for the array-based genotyping, except the missingness threshold was raised to 50%.

2.4.13 Protein-protein interaction and pathway analyses

To identify potential cross-talk among known AD genes and genes suggested by EWASplus, we used web platform GeNets[38] (https://apps.broadinstitute.org/genets) to query a combined list of 28 known AD-associated genes and 123 genes near the top 100 detached CpGs ranked by EWASplus prediction scores.

2.5 Discussion

EWAS has been shown to be a powerful and effective approach to derive associations between methylation changes and phenotypes. EWAS studies of human brain have elucidated additional genes involved in AD[28-31]. To expand our understanding of potential AD-relevant regions in the genome, we developed EWASplus to explore the 97% of CpGs that are not included on the methylation arrays. EWASplus uses an ensemble learning-based computational pipeline to learn relevant features from a large set of potential omics features.

EWASplus is a powerful machine learning method based on disease-specific EWAS results and has some parallels with genotyping imputation strategy used in genetics studies[63]. The fundamental difference is that genomic imputation relies on linkage disequilibrium[64], but DNAm does not share the same degree of physical correlation[33]. In fact, the correlation of methylation levels between two adjacent CpGs decays rapidly with distance[65, 66]. Thus, EWASplus takes an alternative approach by inferring whether a CpG is trait-associated. This is achieved under a supervised classification framework.

EWASplus can effectively identify AD-associated differentially methylated CpGs according to multiple experiments conducted to evaluate its performance. First, using only CpGs on the methylation array, in silico cross-validation revealed high AUC and AUPRC for all six traits. Second, we observed good separation of EWASplus prediction scores between near positive CpGs versus negative CpGs not in the training set (Wilcoxon rank-sum test p-value ranges from 2.82×10^{-99} to 3.64×10^{-16}) in EWASplus prediction scores. For sites not assayed by the methylation array, we found significant enrichment of high-scoring CpGs in genomic regions of functional annotations such as TSS regions and enhancers. Finally, and most importantly, we performed experimental validation using targeted bisulfite sequencing on CpGs not included on the methylation arrays.

Our EWASplus results are notable, in general, for two reasons. First, high-scoring EWASplus CpGs are more likely to be located in regions with functional annotations such as enhancers or promoters. Both of these results are consistent with other work showing that gene regulation is a key facet of many diseases[67]. Second, EWASplus results illustrate how epigenetic “fine mapping” may illuminate disease pathophysiology. For example, in the HOXA locus EWASplus results suggest that epigenetic changes are occurring across the gene cluster in AD rather than one gene-family member.

A key idea of EWASplus is that it bypasses inferring the individual-level DNAm level directly. A similar approach has been used to predict additional trait-associated genetic variants using GWAS and machine learning[12]. Since our goal is to identify disease-associated DNAm CpGs rather than methylation status directly our approach avoids much complexity associated with accounting for the many factors that can influence DNAm CpG status (e.g., age, cell type proportion). This is illustrated by the performance of

EWASplus compared to the modification of Zhang et al.'s method to address disease-association, which it was not originally designed to do, admittedly.

EWASplus results for AD reveal several interesting biological insights. First, we identified a 40 kb region in the homeobox A cluster of genes that are associated with AD, which expands upon the previously described association with individual genes within that cluster (e.g., HOXA3) and AD. Since these are known transcription factors, these findings may suggest important transcriptional regulation occurs in AD or its progression. Second, we find enrichment of kinases—ALPK3, DMPK, MAP3K11, MAP4K1, and TAOK3—in the top EWASplus loci. This finding is particularly relevant for AD given that the pathologic hyper-phosphorylation of tau is a hallmark neuropathologic feature of AD (i.e., neurofibrillary tangles). Of these kinases, only ALPK3 and MAP4K1 were previously suggested to associate with AD[35, 68-70]. DMPK is notable for causing myotonic dystrophy type 1 due to a repeat expansion within an intronic region in carriers that leads to altered gene expression of genes within that region[71]. Interestingly, differential DNAm of MAP4K1 has been associated with AD in human hippocampus[70] and Braak staging (a measure of neurofibrillary tangle pathology)[35] in independent human brain datasets. TAOKs (thousand and one amino acid kinases, also referred to as prostate-derived STE20-like kinases [PSKs]) have been extensively investigated for their ability to phosphorylate MAPT and regulate microtubule assembly[72]; yet, to our knowledge, methylation of TAOK3 has not been previously associated with AD. Finally, from the top 10 EWASplus results (Table 2-2) we found four genes that have intriguing connections with AD or cognitive decline from approaches other than methylation. These genes include DUSP1, PPARD, JUN, and PRKAG2. For example, a PPARD null mouse model shows cognitive impairment[73], and PPARD is highly expressed in the brain[74] and implicated in type 2 diabetes and obesity[75], which are risk factors for AD. In addition, there is experimental evidence to suggest that JUN and PRKAG2 regulate or interact with APP[44, 76], which is of interest in AD given APP is cleaved to beta-amyloid. Thus, these findings from the literature provide complementary support that EWASplus identifies disease-relevant findings and is likely to provide fresh insight into AD.

DNA methylation is tissue-specific. Most of the tests done in this study are conducted on the PFC region of the brain. We focused on PFC for several reasons. First, epigenetic marks are correlated across neocortical regions[77]. Second, cell loss in PFC is relatively less even in people with high neuropathological burden from AD compared to other cortical regions. Third, the majority of available reference human brain transcriptomes and proteomes are from the PFC allowing future work to test predictions of EWASplus using existing data. Despite focusing on PFC, EWASplus performs well on the middle temporal gyrus. Thus, we expect EWASplus to perform well for other tissues because the genome-wide features used are from many different tissue types. From all the tests we performed, we found that the number and level of significant CpGs seem to have a strong impact on the EWASplus performance. Therefore, we are confident that EWASplus will be able to successfully extend the coverage of high quality, well-powered array-based EWAS studies.

Although the EWASplus methodology is general and can be applied to any tissue type, the methylation profiles are tissue-specific, may change with age/environment and demographics. This implies that the trained EWASplus model is only valid for the specific tissue type collected from samples with certain age/environmental profile and demographics. One should exercise caution when trying to extrapolate the results to other tissue types such as blood, or subjects with different age or environmental and demographic profiles. Since the major utility of EWASplus is to expand the coverage of EWAS beyond the array within a specific experimental dataset, this limitation will not hamper the utility of EWASplus.

A potential limitation of EWASplus is the limited number of underlying training datasets and the focus on subjects of recent European descent. Thus, it is of particular importance to expand the number and diversity of additional EWAS data in future work. The underlying methylation data were also from PFC, which is affected relatively late in AD; however, the findings may not generalize to other neocortical regions. Thus, training data from additional relevant brain regions would improve EWASplus models. Likewise, while we started with a large number of potential features, many were from non-neuronal sources, which may limit generalizability to brain tissue. However, as those data are generated, our approach can be easily retrained with those data for improved specificity for brain-cell types from different regions. A strength of this work is that the underlying methylation data were derived from participants enrolled in a population-based study of aging, and there is a wide range of neuropathology findings that reflects the general population rather than a clinic-based ascertainment[78]. We also show a high degree of experimental validation and note that future work could employ targeted bisulfite sequencing[79] or a custom array platform[80] to profile candidate CpGs in a cost-effective and high-throughput manner.

EWASplus does not provide a significant cut-off threshold since it is a supervised classification approach, not a testing-based method. In practice, one can select the threshold empirically by checking whether top CpGs identified by array-based EWAS made the cutoff. Deciding on the number of significant EWAS CpGs to include in training is a tradeoff between the quantity and quality of the training set in EWASplus. Thus, the significance threshold for each EWAS should be decided based on the effect size and sample size of the EWAS. Future work should examine the utility of including different thresholds and use cross-validation to select the desired significance cutoff. For a CpG, no matter how highly ranked by EWASplus, should only be considered as “putative” in terms of trait association unless it can be validated using experimental approaches.

In conclusion, we present EWASplus, a powerful machine learning approach to identify disease-associated CpGs with high reliability. Application of EWASplus to AD highlights important regions and genes that likely contribute to AD pathogenesis, which a valuable addition to the investigation of the epigenetic landscape of AD. In addition, EWASplus is a general approach that may be applied to extend any existing EWAS results obtained using array-based technology, regardless of the trait or phenotypes being studied. We anticipate more exciting findings from its future applications.

Chapter 3 *CASAVA*: A disease category-specific annotation of variants using an ensemble learning framework

3.1 Introduction

Understanding the role of genetic variants in causing complex diseases is a fundamental problem in genetics[81, 82]. Investigators have conducted thousands of genome-wide association studies (GWASs) and identified tens of thousands of loci implicated in human traits and diseases over the past decade[83]. Most of the disease-associated genetic variants lie in the non-coding regions, and many of them are even far away from the nearest protein-coding genes[84]. Thus, delineating the functional implications of these non-coding genetic variants is a significant challenge, requires strategies different from the ones developed to assess coding variants. A possible assumption is that variants in the non-coding regions affect the risk of complex diseases by altering the gene regulation rather than directly affecting protein functions[85]. Currently, some large-scale functional genome projects such as ENCODE and Roadmap Epigenome Mapping Consortium (REMC) have collected massive amounts of sequencing data and thus provided excellent opportunities for annotating non-coding variants[1, 2]. This sequencing-based genome-wide profiling data yields diverse, large-scale genomic or epigenomic features, such as chromatin accessibility, histone modification, transcription factor binding, and gene expression. These features play important roles and could affect the gene regulation process. Many of them have already been utilized as essential sources for functional annotation of non-coding variants[86].

Machine learning has been successfully applied to predict the pathogenicity of genetic variants; however, these methods may not be suitable for prioritizing disease-implicated risk variants due to diverse pathogenicity of complex human diseases and traits. Therefore, it is desirable to develop diverse models to identify disease-specific risk variants. In a recent study, Chen et al. considered the specificities of diseases and presented DIVAN, a method that aims to identify disease-specific risk variants[12].

Although Chen et al. demonstrates the feasibility of using machine learning methods to predict variants in a disease-specific manner, the success of such a strategy hinge upon the availability of sufficient and high-quality training data. But in reality, the number of training risk variants for a specific disease is usually very small, which may lead to inaccurate and unstable models. At this stage, only a few well-studied diseases, such as type 2 diabetes and Coronary Artery disease, have a sufficient number of known disease-associated variants. Hence the applicability of disease-specific variant prediction is very limited. On the other hand, many diseases are related—for example, Alzheimer's disease (AD) and mild cognitive impairment (MCI)—and one disease may be a subtype of another—for example, Late Onset Alzheimer's disease (LOAD) and AD. And many related diseases belong to certain disease categories such as neurodegenerative diseases and autoimmune diseases. These relationships may be explored to help us overcome the problem of insufficient positive training data. In this work, we explore an alternative strategy of finding a middle ground between disease-specific prediction and disease-agnostic prediction. The CASAVA method, or disease CAtegory-Specific Assessment of VAriants, uses disease category information to pool related dis- eases into groups in order to significantly boost the size of the positive training set. CASAVA presents a promising new way to provide both comprehensive and disease-related prediction to sequence variant. Another unique feature of CASAVA is that in order to mitigate computation cost, CASAVA scores are calculated at a 200-bp resolution. That is, genome-wide disease category-specific scores are calculated for every 200-bp bin throughout the human genome. The CASAVA scores for a variant are then taken from the CASAVA scores of the bin that contains the variant. In other words, variants located inside the same 200-bp bin share the same set of CASAVA scores. Despite the reduced resolution, we show that the CASAVA scores provide competitive prediction of disease category-specific risk. The discriminating ability of CASAVA comes from leveraging rich sequencing features and ensemble learning skills effectively. Furthermore, the CASAVA risk scores can be applied to prioritize risk variants in the context of specific diseases and traits.

3.2 Methods

3.2.1 Risk variants for diseases and disease categories

We collected risk variants for specific diseases using the PheGenI database [24]. In order to study the function of variants in non-coding regions, we only retained variants with functional context 'Intron' or 'Intergenic.' For each individual disease, after removing duplications, we sorted them according to P-value and then assigned them to training sets and testing sets in a ratio of 4:1 sequentially from top to bottom in an ordered way.

According to the Medical Subject Headings[87] and PheGenI[88], we used 24 representative disease categories. Each category covers multiple diseases, and one disease may belong to more than one category. Thus, for each category, we combined all training sets of individual diseases belonging to this category in order to constitute the training set for a given disease category. We did the same to obtain the testing set for the disease category and excluded any risk variants in the testing set that are located within 1 kb of any training risk variant.

3.2.2 Constructing control sets of benign variants

Given a set of risk variants, we constructed a corresponding control set of benign variants using a similar strategy as in GWAVA-TSS and DIVAN. We started by downloading all non-coding variants in the 1000 Genomes Project phase 1 release[89]. To minimize the chance that a benign variant would be disease-implicated, we excluded all variants found within 1 kb of any of the variants found in the PheGenI database[88]. Next, we exclude variants with minor allele frequency less than 1% to match the allele frequency range of the risk variants. Finally, we sampled ten times more benign variants than risk variants and required the benign variants to have roughly the same distances to the nearest transcription start sites (TSS) with risk variants (the two empirical distributions of the distances are almost identical). For testing variants, we repeated the sampling procedure ten times.

3.2.3 Processing sequencing features

We adopted the following procedure to process data produced from sequencing-based assays (including the assay for transposase-accessible chromatin using sequencing ATAC-seq, total RNA-seq, and whole genome bisulfite sequencing WGBS) into features to be used in our machine learning models. We first downloaded mapped reads from the ENCODE and the ROADMAP project[1, 86]. For mapped reads using hg38 assembly, we applied genomic coordinates conversion from hg38 to hg19. Most of the experiments in ENCODE contained biological replicates, and we merged read counts from different technical replicates into a single feature. After processing, we got 66, 243 and 255 features of ATAC-seq, RNA-seq and WGBS, respectively. We also downloaded 355 processed datasets of gene expression (in transcript per million (TPM) formats). For each genetic variant, we calculated the expression of its nearest gene in different tissues / cell-lines. Additionally, we inherited the 1806 features used in DIVAN. In total, we amassed 2,725 genome-wide features, which can be roughly divided into five groups: open chromatin, histone modification, TF binding, gene expression and DNA methylation.

To simplify the calculation, we divided the entire genome into 200-bp bins and calculated the normalized mapped read counts for each bin. We stored the resulting features in a 15,685,849 by 2,725 matrix. For this matrix, each row represents a 200-bp bin, and each column represents a feature. For a genetic variant, we first found which bin the variant fell into, then retrieved the corresponding feature values.

3.2.4 Ensemble learning for class imbalance problem

To train CASAVA models, we adopted an ensemble learning strategy by combining the gradient boosting regression tree and a bagging technique[57, 90]. The input data are labeled training data (risk variants and benign variants along with their weights). Each variant is represented by 2,725 features. For CASAVA, the weight of each variant was set as default value 1. For each training round, we took all risk variants and randomly sampled a subset of benign variants such that risk and benign variants had an approximately equal sum of weights[90]. Based on XGBoost, we trained a gradient boosting regression tree classifier using these selected variants[57]. We repeated the under-sampling and training process a number of times (e.g., 100 times) and took their average as our final model. We trained a total of 24 models for disease categories (CASAVA models).

To achieve the best performance, we made several adjustments to the algorithm adopted by DIVAN[12] and GWAVA[91]. First, adaptively using boosting trees (instead of a single tree like GWAVA) provided enough model capacity to deal with different complex diseases. Second, to prevent the boosting trees from overfitting, we used under-sampling 100 times in CASAVA, rather than just 20 in DIVAN. And this specific bagging technique relieved the class imbalance problem in our data and alleviated the need for parameter-tuning.

3.2.5 Genomic properties of CASAVA score

We downloaded all the genetic variants from the 1000 Genomes Phase 3 release[89] and predicted these variants using CASAVA scores. According to the Ensemble Variant Effect Predictor[92], we assigned each variant to one of the following genomic contexts: ‘promoter,’ ‘exon,’ ‘intron,’ ‘intergenic’ and ‘1 to 5 kb.’ The term ‘1 to 5 kb’ indicates the regions located 1000-bp to 5000-bp upstream of the transcription start sites (TSS). To emphasize the importance of the enhancer region, we assigned the genomic context of a variant the label ‘enhancer’ if it located in the FANTOM enhancer region[93]. Please note we used these annotated enhancers for the purposes of illustrations without considering the cell-type specificity. To further clarify, ‘intergenic’ indicates intergenic regions excluding the enhancer regions.

Next, we binned the variants according to the quantiles of CASAVA scores. Within each bin, we calculated the proportion of variants with different genomic contexts. Given a disease category, variants with the top 10% CASAVA scores were denoted by high-score variants. Next, we performed chi-square test (a two-by-two table) to see whether variants with a specific genomic context (e.g., enhancer regions) were over- or under-represented among these high-score variants. We also made a normalized version to better reflect the relative composition of these genomic components. We calculated the proportion of variants with a specific genomic context after normalizing by the total number of variants located in regions with the genomic context.

3.2.6 Applying CASAVA to disease-specific risk prediction

We leveraged CASAVA to predict disease-specific risk variants. Given a specific disease, we first identified its corresponding disease category/categories using MeSH and took the average of its category scores as an approximation of the disease-specific score. For example, the Hodgkin disease belongs to three different disease categories: hemic and lymphatic disease, immune system disease, and neoplasm. We took the average CASAVA scores of hemic and lymphatic diseases, immune system diseases and neoplasms as an approximation of the score of the Hodgkin disease.

We tested the CASAVA's ability to predict disease-specific risk for variants. Some variants are associated with multiple diseases. In order to best maintain independence between the training set and the testing set, we excluded risk variants in the testing set that are located within 1 kb of any training risk variant. We benchmarked the results on 89 diseases which had more than 50 known disease-associated variants in its training set and at least 10 risk variants in its testing set. Besides, the trained CASAVA models also used risk-training variants of these diseases. Thus, we merely evaluated the success of this approach on diseases that the training set had seen before. We did a simulation study to mimic a scenario in which there is no training data at the disease-level. Given a specific disease among the 89 diseases, we used all its associated variants as testing variants. We excluded the corresponding training variants, retrained the CASAVA models, and reevaluated the approximation approach.

3.2.7 Applying transfer learning to disease-specific risk prediction

We leveraged information from related-diseases to boost the performance of disease-specific prediction using the transfer learning technique[94]. For a specific disease, we denoted its training variants by 'disease-specific training variants' and used the training variants belonging to other diseases in this disease category as ‘disease category-specific training variants’. In order not to over-estimate the model performance, we excluded disease category-specific training variants which overlap with any disease-specific training variant or testing variant. After giving more weight to disease-specific training variants (e.g., weight=5), we combined them with disease category-specific training variants, and trained transfer learning models using the previous ensemble learning method.

3.2.8 Comparison with commonly used scoring methods

We compared CASAVA with ten existing functional impact prediction methods: CADD[8], DANN[95], GWAVA[9], FATHMM-MKL[91], GenoCanyon[58], deltaSVM[96], Eigen[97], DIVAN[12], LINSIGHT[98] and PAFA[99]. Though these methods utilized different hypotheses and techniques, they are all reported to be informative of risks of complex diseases. For each method, we downloaded their pre-computed scores and scored the testing variants. For GWAVA, we used the unmatched, TSS-matched, and region-matched scores. Due to the problem setting, we only considered non-coding scores of FATHMM-MKL. For deltaSVM, we downloaded the saved model, which was trained from GM12878 DNA hypersensitivity sites and scored the variants. For DIVAN, to make a fair comparison, we used the same training pipeline as in the original study and retrained it on the specific diseases we tested.

3.2.9 Performance evaluation

The receiver operating characteristics curve (ROC) is a typical graphical plot that illustrates the classification ability of a binary classifier system[100]. We also considered the precision-recall curve due to the imbalance between risk and benign variants[101]. We used both the area under the ROC (AUC) and area under the precision-recall curve (AUPR) to assess the prediction performance for each task, and calculated the AUC and AUPR values using the ROCR package[102]. We first evaluated the performance of each method by five-fold cross-validation on training variants. Then, we estimated the performance of each method using independent testing variants. To eliminate bias, we repeated the sampling procedure ten times, given a set of risk variants for testing. Each time, we used a different set of benign variants, and calculated the average AUC and AUPR values across these ten repetitions.

3.2.10 Case study for immune system diseases

We downloaded all the genetic variants from the 1000 Genomes Phase 3 release. We predicted each variant with 24 CASAVA scores. First, we selected variants whose immune system disease scores are the highest among the 24 scores. We excluded variants located within 10 kb of any training variant. For each variant, among its 24 CASAVA scores, we calculated the ratio of its second highest score divided by its highest score. Next, we sorted the variants according to the ratio in ascending order, as a lower ratio shows better specificity in terms of disease category classification. We applied the threshold of 0.7 to select variants for further validation. For all candidate loci, we performed batch query in SNPnexus[103] for their known disease-phenotype association. SNPnexus is an interface of a collection of SNP functional annotation databases that can be used for querying the validated disease information of the submitted SNPs in GAD[104], COSMIC[105], and CinVar[106] databases; to fit the aims of our ensemble classifier, we focused on the query results of the GAD database[104] since the annotation of each association contains both disease class and disease name. We found a few validated variant-disease associations with annotated category 'IMMUNE' in the GAD database, along with the nearest genes of the variants. For the purposes of illustration, we took two genes, MHC2TA and IKZF1, to show the usefulness of CASAVA scores.

3.2.11 Exploring informative features in CASAVA

For a gradient boosting regression tree model, the 'relative importance' of a feature is in percentage format, indicating how much the feature contributes to constructing the model. We computed the relative importance of each feature using the XGBoost[57] R package and used the average relative importance of the 100 base models as the value for the CASAVA model. Given a group of features, we used the sum of their relative importance (of each feature) as the relative importance of this group. Then we calculated the relative importance of feature groups related to histone modifications, open chromatin, TF binding, gene expression, and DNA methylation.

Next, we combined all risk and benign variants from 24 CASAVA training sets and removed duplications. Given a sequencing feature, we extracted the counts from upstream 4000 bp to downstream 4000 bp of each genetic variant and formulated the count in 200-bp bin format. For each variant, we got 8200 / 200=41 numbers in order and transferred the counts into log2 scale. At each of the 41 relative positions of variants, we calculated the average counts for all risk/benign variants and drew line plots. For the purposes of illustration, we used DNase, H3K9me3, H3K4me1 and H3K27ac of A549 cell line.

3.3 Results

3.3.1 Overview of CASAVA

The goal of CASAVA is to provide a comprehensive prediction of disease risk in 24 disease categories for any non-coding variant in the genome. The result is a 24-component vector: each component is a continuous score ranging from 0 (minimum risk) to 1 (maximum risk) to indicate risk of predisposing to diseases in one of the 24 disease categories (Figure 3-1). To achieve this, we designed an ensemble machine learning strategy and implemented a two-step procedure. First, we calculate a set of CASAVA scores for every 200-bp bin throughout the genome using the trained models. Next, we assign the CASAVA scores for the bin to all the variants located inside the bin. In other words, the resolution of the CASAVA scores is 200 bp.

In the present study, we focus on variants located in the non-coding part of the genome. Given a disease category, we first collect relevant non-coding risk variants (located in intron or intergenic only) from PheGenI[88] using significance level threshold of 10^{-4} (Figure 3-1a). We next select corresponding control sets of benign variants from the 1000 Genomes project for each disease category. In the meantime, we collect, curate, and process a large set (2,725) of genome-wide profiles to be used as features in the classification model (Figure 3-1b). For many complex diseases, there may exist multiple distinct routes in disease pathogenicity. For example, many diseases have subtypes. For each of these subtypes, a unique biological mechanism may be involved. And the omics profiles of these subtypes may be different. Hence, for a single disease, there may exist multiple omics patterns around its risk variants. We are hoping that each of these patterns can be captured by one or few of the base learners in the ensemble model. To account for the heterogeneity in the disease pathology, we opt for an ensemble learning strategy, which is capable of recognizing multiple omics profiles in making the prediction. For base learners, we choose boosting trees with the bagging technique (Figure 3-1c).

In the end, CASAVA trains an ensemble classifier for each of the 24 broad disease categories (Figure 3-1d) and applies the trained model genome-wide to calculate disease category- specific scores. These scores can be used to assess disease risks in the most common disease categories. To make CASAVA easily accessible, we build a web portal to allow easy browsing and querying of CASAVA scores along with visualization (http://zhanglabtools.org/CASAVA).

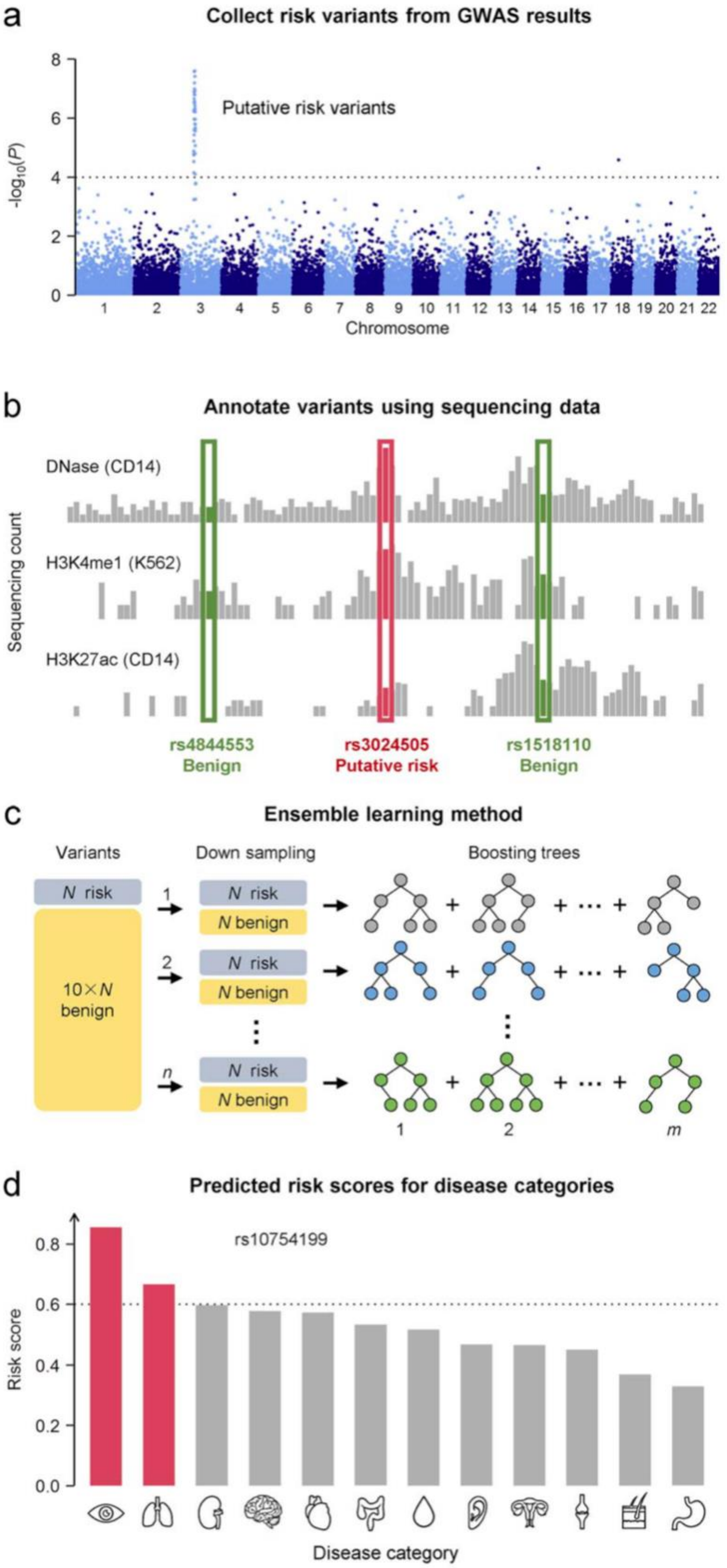

a
Collect risk variants from GWAS results
Putative risk variants
-log10(P)
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome
b
Annotate variants using sequencing data
Sequencing count
DNase (CD14)
H3K4me1 (K562)
H3K27ac (CD14)
rs4844553
Benign
rs3024505
Putative risk
rs1518110
Benign
c
Ensemble learning method
Variants
N risk
10× N benign
Down sampling
1
2
n
N risk
N benign
Boosting trees
1
2
m
d
Predicted risk scores for disease categories
rs10754199
Risk score
0.8
0.6
0.4
0.2
0.0
Disease category

Figure 3-1 Working pipeline of CASAVA.

(a) For each disease category, CASAVA collects known risk variants from existing genome-wide association studies (GWASs) as training data. (b) CASAVA uses genome- wide genomics and epigenomics profiling data as features in its machine learning models. (c) CASAVA applies bagging and boosting techniques to build a classification model for each disease category. (d) CASAVA produces disease category-specific risk prediction for non-coding genetic variants.

3.3.2 Disease categories

For disease categories, we use those defined by the Medical Subject Headings (MeSH) related to ‘diseases’ or ‘psychiatry and psychology’. The same definition was also used by PheGenI[88]. Next, we exclude the parasitic disease category due to an insufficient number of variants (less than 100) associated with its member diseases. We also exclude five disease categories that are unlikely to have a strong genetics component: animal diseases, chemically-induced disorders, disorders of environmental origin, occupational diseases, and wounds and injuries. For the remaining 24 categories, using the aforementioned significance threshold, the numbers of their associated non-coding risk variants cataloged by PheGenI range from 137 to 8,065 with a median of 1,337 (Table 3-1). The total number of non-coding variants for the 24 disease categories is 29,233. According to PheGenI, these variants are associated with 484 individual diseases. The number of associated variants of these diseases ranges from 1 to 2,995, with a median of 15.

Ind ex	Disease category	Abbreviation in paper	# Risk variants (train)	# Risk variants (test)
1	Bacterial Infections and Mycoses	Bacterial and Mycoses	172	35
2	Behavior and Behavior Mechanisms	Behavior Mechanisms	2232	468
3	Behavioral Disciplines and Activities	Behavior Disciplines	591	133
4	Cardiovascular Diseases	Cardiovascular	3802	810
5	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Congenital and Hereditary	671	115
6	Digestive System Diseases	Digestive System	1452	240
7	Endocrine System Diseases	Endocrine System	1237	252
8	Eye diseases	Eye	766	152
9	Female Urogenital Diseases and Pregnancy Complications	Female Urogenital	651	125
10	Hemic and Lymphatic Diseases	Hemic and Lymphatic	370	72
11	Immune System Diseases	Immune System	1915	386
12	Male Urogenital Diseases	Male Urogenital	947	197
13	Mental Disorders	Mental Disorders	6574	1405
14	Musculoskeletal Diseases	Musculoskeletal	622	125
15	Neoplasms	Neoplasms	2819	562
16	Nervous System Diseases	Nervous System	2640	593
17	Nutritional and Metabolic Diseases	Nutritional and Metabolic	1730	389
18	Otorhinolaryngologic Diseases	Otorhinolaryngo logic	114	23
19	Pathological Conditions, Signs and Symptoms	Pathological Symptoms	1944	408
20	Psychological Phenomena	Psychological Phenomena	881	197
21	Respiratory Tract Diseases	Respiratory Tract	1575	316
22	Skin and Connective Tissue Diseases	Skin and Connective	1695	304
23	Stomatognathic Diseases	Stomatognathic	551	92
24	Virus Diseases	Virus	120	23

Table 3-1 Disease categories of CASAVA.

3.3.3 Predicting disease category-specific risk variants

To evaluate the performance of CASAVA in terms of predicting disease category risk, we first conducted a five-fold cross-validation study, comparing CASAVA with nine scoring methods that provide prediction scores genome-wide: CADD[8], DANN [95], GWAVA[9], GenoCanyon[58], FATHMM-MKL[107], deltaSVM[96], Eigen[97], LINSIGHT[98] and PAFA[99]. We found that overall CASAVA performed the best, followed by PAFA and GWAVA in terms of AUC. We next conducted a follow up study using independent testing sets; CASAVA again achieved the best performance among all methods in terms of AUC and AUPR. Compared to scores from commonly used methods, CASAVA improved the AUC by at least 0.05 for 17 out of the 24 (70.8%) categories and lifted the AUPR by at least 0.05 for 11 out of the 24 (45.8%) categories. Yet, the performance varied tremendously across different tasks. For all of the 24 disease categories, the AUC from CASAVA falls in the range of 0.62–0.78 with a median of 0.68, and the AUPR from CASAVA falls in the range of 0.12– 0.37 with a median of 0.18. For some categories, such as eye diseases, even for its closest competitors, CASAVA's advantage is rather significant (AUC: 0.78 versus 0.62 (Figure 3-2b); AUPR: 0.35 versus 0.14 (Figure 3-2c)). Overall, CASAVA performs the best among all methods we compared in terms of AUC and AUPR value.

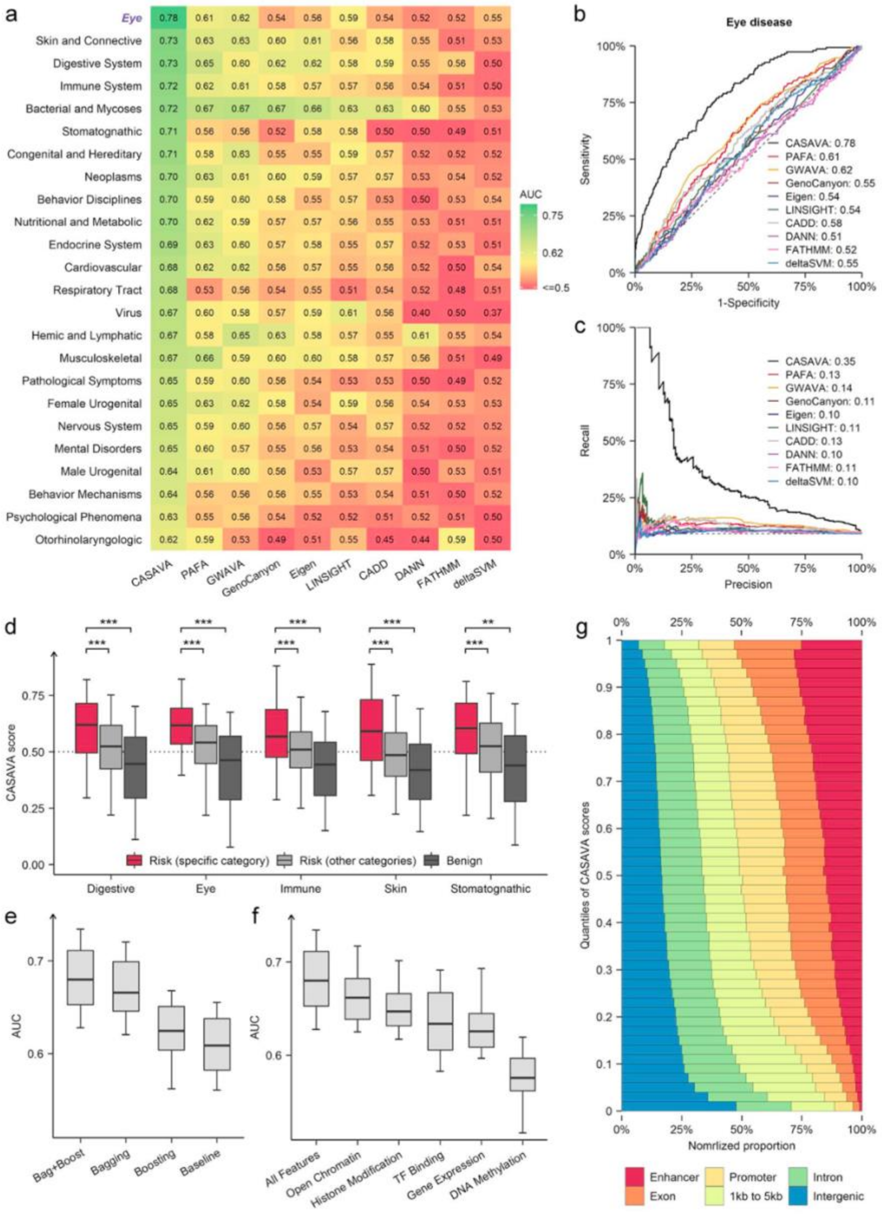


Figure 3-2 Performance evaluation for disease category-specific risk prediction.

(a) Heatmap of AUCs of different methods for 24 disease categories. (b) ROC of different scoring methods for eye diseases. (c) PRC of different scoring methods for eye diseases. (d) Side-by-side boxplots of CASAVA scores comparing three groups of variants namely 'variants associated with diseases belong to the specific disease category,' 'variants associated with diseases belong to other diseases categories' and 'benign variants'. P-value was calculated using the one-tailed Wilcoxon rank-sum test. $* 1 \times 10^{-8} < P \leq 5 \times 10^{-2}$; $** 1 \times 10^{-16} < P \leq 1 \times 10^{-8}$; $*** P \leq 1 \times 10^{-16}$. All boxplot whiskers show 95th/5th percentile. (e) Boxplots of 24 AUC values (one for each disease category) showing difference across various ensemble learning techniques. (f) Side-by-side boxplots of 24 AUC values (one for each disease category) illustrating different level of informativeness across five types of features. 'All features' refers to using all five groups of features. (g) Proportion of local genomic annotation types (e.g., promoters, enhancers) for each CASAVA score bin, after first normalizing by the total number of variants observed in that genomic annotation types. Here we use CASAVA scores for the eye diseases category as an example.

3.3.4 Disease category-specificity in CASAVA scores

All existing methods, except for DIVAN, produce a single score for each variant to represent its pathogenicity. As expected, when comparing known (identified by GWASs) disease-associated variants with benign variants, these methods return higher scores (indicating pathogenicity) for the former (Figure 3-2a). In contrast, CASAVA generates 24 scores for each variant, one for each disease category. For any given disease- associated variant, we want to answer the following two questions: first, does its disease category-matching CASAVA score tend to be higher than that of benign variants? Second, does its disease category-matching CASAVA score tends to be higher than the other 23 unmatching CASAVA scores? For the first question, we found that specific CASAVA scores (from the corresponding disease) of risk variants are significantly higher (one-tailed Wilcox rank-sum test, $p < 0.05$) than those of benign variants (Figure 3-2d) in all 24 disease categories. For the second question, we found in 17 out of the 24 categories (70.8%) that the CASAVA scores of risk variants from the matching disease category are significantly higher (one-tailed Wilcox rank-sum test, $p < 0.05$) than their CASAVA scores from the other 23 disease categories combined (Figure 3-2d). These results demonstrated the disease-category specificity in CASAVA score.

3.3.5 Benefits of using various ensemble learning techniques

The superior performance of CASAVA can be traced back to the key techniques we adopted, including the use of tree-based ensemble models, bagging, and boosting trees. We have showed that applying these techniques indeed made a difference for classification and found that training a single decision tree without ensemble learning produced rather poor results (Figure 3-2e; average AUC = 0.615). Incorporating boosting trees lifted average AUC to 0.637 (one-tailed paired t-test, $p = 8 \times 10^{-9}$). With down-sampling, bagging a series of decision trees further lifted average AUC to 0.683 (one-tailed paired t-test, $P = 2 \times 10^{-11}$). Compared to a single decision tree, using boosting trees with bagging technique improved the AUC values by 0.08 on average (from 0.615 to 0.697), a remarkable performance boost (one-tailed paired t-test, $p = 5 \times 10^{-13}$). Besides, for predicting the risk of disease categories, the ensemble learning algorithm achieved higher AUC and AUPR than traditional machine learning methods like random forest and logistic regression (Figure 3-3).

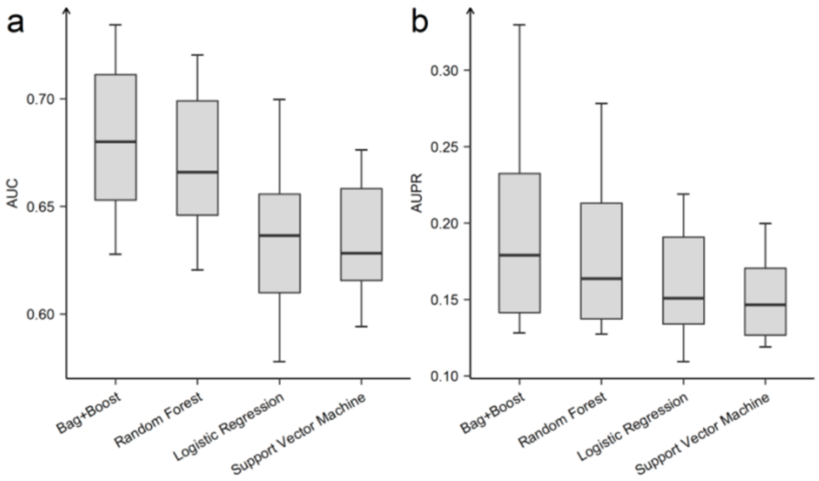


Figure 3-3 Performance evaluation for disease-category risk prediction using different machine learning methods.

a) Side-by-side boxplots of 24 AUC values (one for each disease category) using different machine learning methods. b) Side-by-side boxplots of 24 AUPR values using different machine learning methods.

3.3.6 Contributions from different group of features

The current version of CASAVA utilized 2,725 features. These features can be divided into five groups: open chromatin, transcription factor (TF) binding, histone modification, DNA methylation and gene expression. A natural question is whether every feature group contributes to the success of CASAVA. To answer this question, we did the following experiment. For each of the 24 disease categories, we took turns to only use features from a single feature group (such as the histone modification or TF binding group) to train a classification model and test its performance using independent testing sets. All models achieved significantly higher AUC and AUPR values than random guess, indicating the usefulness of every single group of features (Figure 3-2f).

Features related to histone modification can be divided into two subsets: active (or open) chromatin such as H3K4me3 and H3K27ac, and repressive (or closed) chromatin such as H3K9me3 and H3K27me3 (Table 3-2). Most of the existing variant prediction methods only focus on the uses of open chromatin marks. But we found that for all 24 disease categories, only using features with active or repressive effects leads to average AUC 0.644 and 0.638, respectively. When combined together, we got an average AUC of 0.650, which confirms the usefulness of both subgroups of features. Taken together, our results indicated that closed chromatin marks contributed almost the same as open chromatin marks. And the performance of CASAVA is slightly better when combined both types of histone marks.

	AUC			AUPR		
Disease category	**Histone modification**	**Open chromatin**	**Close chromatin**	**Histone modification**	**Open chromatin**	**Close chromatin**
Bacterial Infections and Mycoses	0.69	0.71	0.66	0.31	0.32	0.23
Behavior and Behavior Mechanisms	0.62	0.62	0.61	0.12	0.12	0.12
Behavioral Disciplines and Activities	0.66	0.64	0.61	0.15	0.15	0.13
Cardiovascular Diseases	0.67	0.66	0.64	0.14	0.14	0.14
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	0.65	0.62	0.64	0.16	0.16	0.16
Digestive System Diseases	0.70	0.70	0.69	0.22	0.23	0.19
Endocrine System Diseases	0.65	0.64	0.64	0.16	0.15	0.15
Eye diseases	0.69	0.69	0.66	0.19	0.19	0.17
Female Urogenital Diseases and Pregnancy Complications	0.63	0.62	0.63	0.15	0.14	0.14
Hemic and Lymphatic Diseases	0.65	0.66	0.64	0.19	0.20	0.16
Immune System Diseases	0.70	0.70	0.70	0.24	0.23	0.21
Male Urogenital Diseases	0.62	0.60	0.61	0.13	0.11	0.13
Mental Disorders	0.63	0.63	0.62	0.13	0.13	0.12
Musculoskeletal Diseases	0.63	0.63	0.64	0.16	0.16	0.15
Neoplasms	0.66	0.66	0.65	0.15	0.15	0.15
Nervous System Diseases	0.63	0.63	0.63	0.14	0.13	0.13
Nutritional and Metabolic Diseases	0.67	0.67	0.64	0.15	0.15	0.13
Otorhinolaryngologic Diseases	0.57	0.56	0.65	0.14	0.12	0.15
Pathological Conditions, Signs and Symptoms	0.64	0.64	0.62	0.13	0.14	0.12
Psychological Phenomena and Processes	0.63	0.63	0.59	0.13	0.13	0.11
Respiratory Tract Diseases	0.62	0.62	0.62	0.13	0.13	0.13
Skin and Connective Tissue Diseases	0.71	0.71	0.68	0.26	0.26	0.22
Stomatognathic Diseases	0.66	0.65	0.63	0.19	0.17	0.16
Virus Diseases	0.64	0.60	0.61	0.14	0.14	0.13

Table 3-2 Performance evaluation of using different groups of histone modification features.

3.3.7 Genome-wide pattern of CASAVA scores

Once all CASAVA scores are derived, it is of interest to explore the distribution of these scores, especially those top scores for each disease category. This may shed light on how genetic variants contribute to disease pathogenesis. For example, we found that for genomic regions with high CASAVA scores for eye diseases, the enhancer regions are significantly over-represented (Figure 3-2g, Chi-squared test, $p < 2.2 \times 10^{-16}$). In contrast, intergenic regions (not in enhancer regions) are depleted (chi-squared test, $p < 2.2 \times 10^{-16}$).). Such a pattern is observed for almost all the 24 disease categories. Our finding is consistent with the notion that most GWAS variants are likely disruptive of transcription regulation of genes critical for the pathogenesis of the disease[108].

3.3.8 Results on testing sets

To mimic the scenario of different testing sets, we performed the following three experiments to compare performance of CASAVA with commonly used scoring methods.

In the first experiment, since all the risk variants stored in PheGenI came from two databases—NHGRI GWAS catalog (NHGRI)[83] and dbGaP[109], we treat risk variants from one source as the training set and risk variants from the other source as the testing set and vice versa. In the second experiment, we divide all the risk variants into two separate groups according to which chromosome they belong. One group consist of all the odd number chromosomes plus chromosome X, and another group consist of all the even number chromosomes and vice versa. In the third experiment, we split the risk variants according to the magnitude of statistical significance. Variants with association P-value lower than a threshold are assigned to the training set and the rest are assigned to the testing set and vice versa. In all three experiments, we found that CASAVA achieves the best performance overall.

3.3.9 Utility of CASAVA scores on disease-specific risk prediction

The goal of CASAVA is to provide disease category-level prediction. Having achieved that, an interesting follow up question is whether the CASAVA scores can also be leveraged for prediction at the individual disease level. Unlike DIVAN, which relies on disease-specific training data, CASAVA scores are trained by aggregating variants from all diseases belonging to the same disease category. Therefore, we hypothesized that CASAVA scores may be particularly informative when disease-specific variants needed for training are scarce or not available at all, which is the case for the majority of complex diseases. We believe using CASAVA scores (for disease category) as surrogate to predict the risk of individual disease is feasible, because, for many disease categories, the same genomic variants have been found to be associated with multiple diseases (Figure 3-4a)[110, 111]. In spirit, our strategy is reminiscent of the transfer learning idea that has proved surprisingly effective in many machine learning applications.

The 24 disease categories include 484 individual diseases with at least one associated non-coding risk variants. The numbers of such variants range from 1 to 2,995 with a median of 15. To get relatively robust results, we used 89 diseases to evaluate the performance of CASAVA at the individual disease-level (Methods); that is, we used CASAVA disease category-specific scores to predict disease-specific risk. We again found that overall CASAVA still achieve the best performance, with an average AUC of 0.692, compared to average AUC of 0.647 for DIVAN and average AUC of 0.607 for PAFA (Figure 3-4b). In terms of AUC, CASAVA achieved the best performance in 59 out of the 89 diseases (66.3%). Furthermore, CASAVA improved the AUC by at least 0.05 in 21 diseases (23.6%) and lifted the AUPR by at least 0.05 in 21 diseases. Yet the prediction performance varied substantially across different diseases. For all 89 diseases, CASAVA produces AUC values in the range of 0.52–0.90 with a median of 0.68, and the AUPR values resulting from CASAVA fall in the range of 0.10–0.58 with a median of 0.17. For comparison, we also trained disease- specific models for these 89 diseases using our ensemble learning framework. CASAVA presented results that are comparable to disease-specific models on the 89 diseases in terms of AUC (Figure 3-4c, Pearson correlation = 0.79).

In the above result, we saw that the performance of CASAVA is still better than DIVAN. This is because that DIVAN designed for disease-specific risk prediction limits its application to only diseases with large number of known disease-specific variants (needed for training). CASAVA overcomes this limitation by focusing instead on the 24 major disease category which gives much larger training set. In addition to disease category-specific risk prediction, a secondary, and admittedly suboptimal application of CASAVA is to predict disease-specific risk, simply by borrowing disease category-specific CASAVA scores from the disease category that contains the particular disease. For the majority of diseases where only a small number (less than 20) of known disease-associated variants in known, CASAVA have distinct advantage.

Next, we conducted a simulation study to mimic a scenario in which there is no training data available at the individual disease level. Given a disease, we treat all its known variants associated with it as testing data. We removed these variants from training sets, re-trained CASAVA, and evaluated its performance. Surprisingly, this seemingly simple-minded approach again achieved remarkably better results than existing methods in terms of AUC and AUPR (Figure 3-4d). Using the same 89 diseases, in terms of AUC, CASAVA achieved the best performance in 81 out of the 89 diseases (91.0%). Moreover, CASAVA improved the AUC by at least 0.05 in 47 diseases (52.8%) and lifted the AUPR by at least 0.05 for 17 diseases (19.1%).

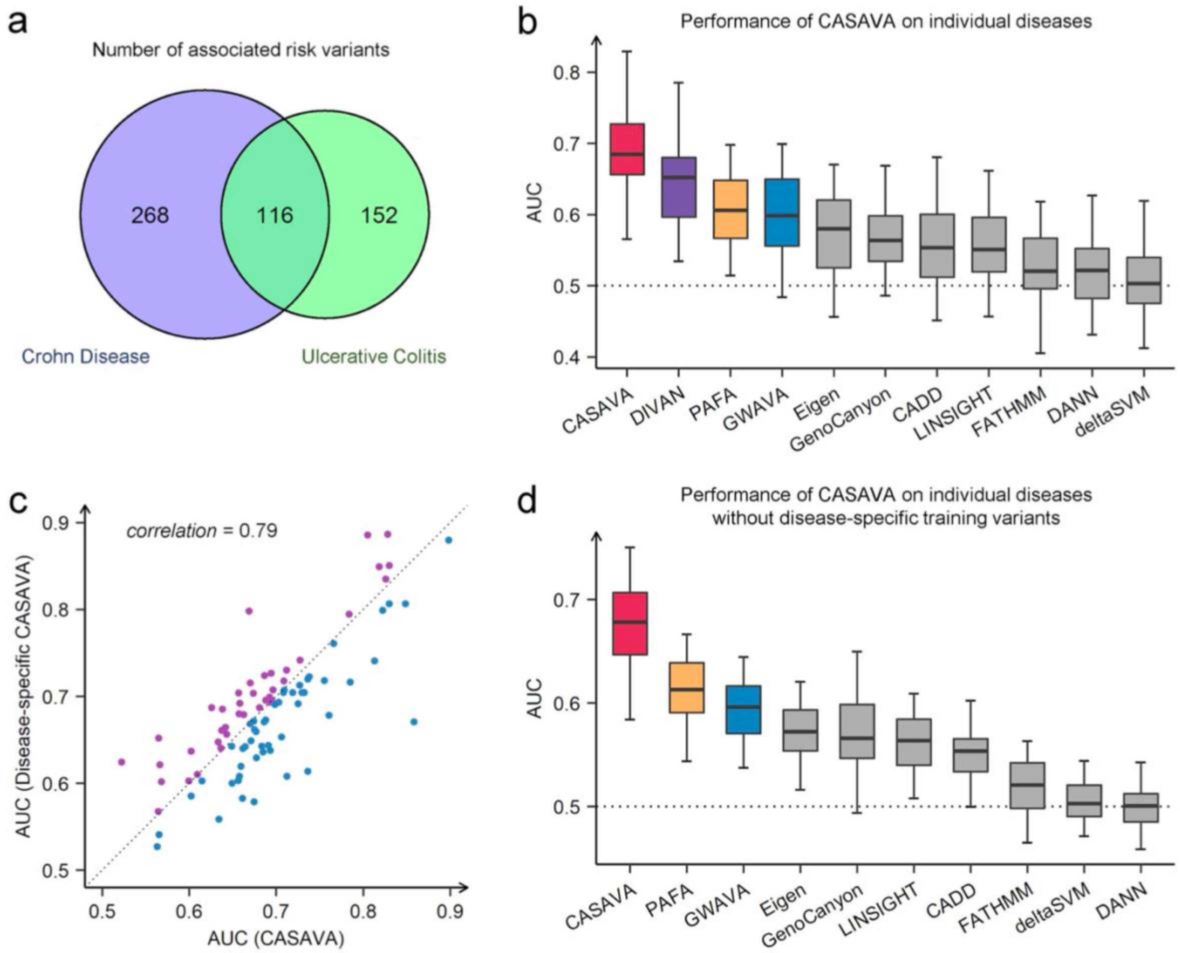


Figure 3-4 Performance evaluation for disease-specific risk prediction.

(a) Venn's diagram for known risk variants that belong to two digestive system diseases. (b) Side-by-side boxplots of 89 AUC values (one for each disease) comparing performance between CASAVA and ten different variant prediction methods. Some methods have multiple scores, and we only use the score with the highest average AUC values. GWAVA score is in fact GWAVA TSS-matched score, and Eigen score is actually Eigen- PC score. (c) Scatter plot comparing AUC values obtained using two different methods: regular (disease category-specific) CASAVA and the disease-specific version of CASAVA (apply the same ensemble learning framework to each of the 89 diseases). Each point represents one of the 89 diseases. We use Pearson's correlation coefficients as the correlation measure. Purple and blue represent the condition where one method outperforms the other one. (d) Side-by-side boxplots of 89 AUC values (one for each disease) comparing performance between CASAVA and nine different variant prediction methods, assuming no disease-specific training data is available. Here disease-specific training variants were excluded when training each of the CASAVA models.

3.3.10 Applying transfer learning to improve disease-specific risk prediction

Previously, we demonstrated the utility of directly using CASAVA scores designed to predict disease-category risk for diseases belonging to the disease category. Despite the decent results of this strategy, we felt that a better approach would be to use both variants associated with the specific disease and variants associated with similar diseases in the same disease category. This strategy is particularly important when disease-specific variants are scarce.

To accomplish this, we designed an instance-based transfer learning approach named TrCASAVA, which includes both individual disease level variants as well as disease category-level variants in the training set. TrCASAVA applies higher weights to disease-specific variants to prioritize variants at the individual disease level.

Compared to the disease-specific models, our results showed that TrCASAVA improves performance in 64 out of the 89 diseases (71.9%) in terms of AUC (Figure 3-5a). On average, TrCASAVA lifted the AUC value by 0.013 (Figure 3-5b, one-tailed paired t-test, $p = 4 \times 10^{-4}$) and the AUPR value by 0.015 (one-tailed paired t-test, $p = 6 \times 10^{-4}$). Compared to CASAVA, TrCASAVA also achieved higher AUC values on 54 out of the 89 diseases (60.7%), which was possibly due to utilizing disease specificities (Figure 3-5c). On average, TrCASAVA lifted the AUC value by 0.005 (Figure 3-5d, one-tailed paired t-test, $p = 0.04$) and the AUPR value by 0.009 (one-tailed paired t-test, $p = 0.02$).

We also did an ablation study assuming that only a small number of disease-specific training variants were available and performed experiments on 57 diseases with more than 100 disease-associated variants in the training set. Chen et al. showed that the performance of a disease-specific variant prediction model is highly dependent on the size of the training set. Using few disease-specific training variants led to rather poor results (Figure 3-5e). Under the scenario of an extremely small training set, TrCASAVA is likely to significantly improve the prediction results (Figure 3-5e). For example, if we only included 1/8 of the disease-specific variants for training, we got an average AUC value 0.64 while TrCASAVA lifted the average AUC value by 0.05 (one-tailed paired t-test, $P = 4 \times 10^{-14}$). Put together, we concluded that the predictions of TrCASAVA are more robust than those of the disease-specific models trained on a small number of variants.

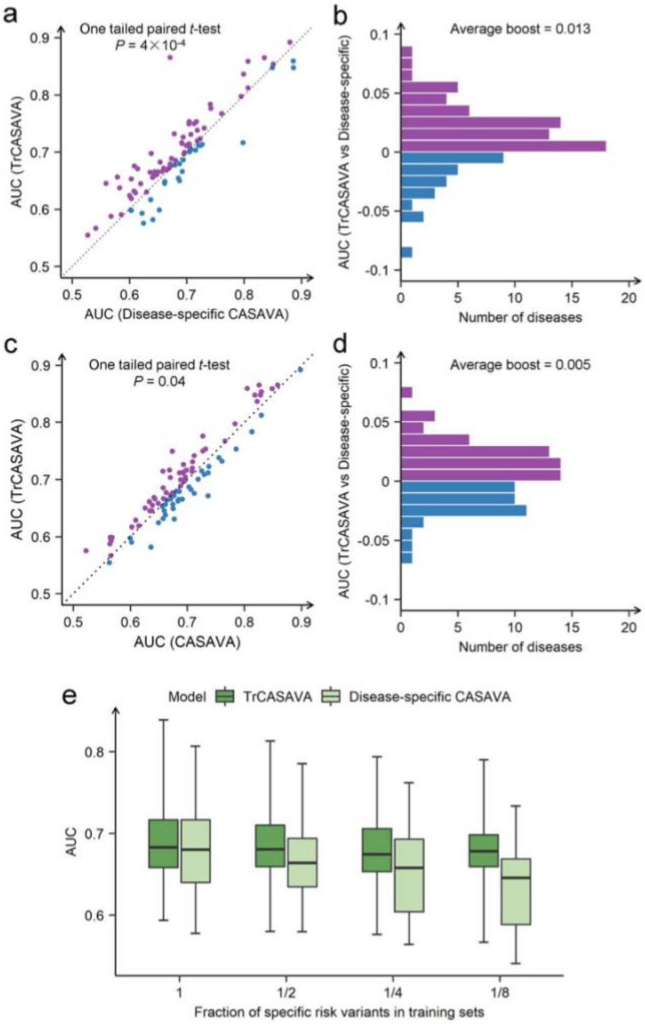


Figure 3-5 Performance of TrCASAVA for disease-specific risk prediction.

(a) Scatter plots comparing AUC values obtained using two different methods: TrCASAVA and the disease-specific version of CASAVA (apply the same ensemble learning framework to each of the 89 diseases). Each point represents one of the 89 diseases. P-value is calculated using the one-tailed paired t-test. Purple and blue represent the condition where one method outperforms the other one. (b) Histogram of AUC differences between TrCASAVA and the disease-specific version of CASAVA (apply the same ensemble learning framework to each of the 89 diseases). (c) Scatter plot comparing AUC values obtained using TrCASAVA and CASAVA. (d) Histogram of AUC differences between TrCASAVA and CASAVA. (e) Side-by-side boxplots of 57 AUCs (one for each disease) comparing performance of TrCASAVA and the disease-specific version of CASAVA with varying fraction of disease-specific variants in the training set. All boxplot whiskers show 95th/5th percentile.

3.3.11 Case study: MHC2TA and IKZF1 for immune system diseases

The relationship between MHC2TA and immune system diseases has long been noticed and documented in the literature[112]. As reported, polymorphisms in and around the MHC2TA gene lead to differential MHC molecule expression and are associated with susceptibility to diseases with inflammatory components[113]. For example, the variant rs3087456, located in the promoter region of MHC2TA gene, has been shown to increase susceptibility to rheumatoid arthritis and multiple sclerosis[113]. The CASAVA scores seem to agree with this fact. In the gene body region of MHC2TA, the average score of immune system diseases is the highest among all 24 disease categories (Figure 3-6a, Figure 3-6b). We further explored the CASAVA scores of 2144 variants located within the gene body as well as 5-kb flanking regions of MHC2TA. The scores of immune system diseases achieved the highest and the second highest in 1012 (48%) and 861 (41%) out of the 2144 variants, respectively (Figure 3-6c). And for 89 variants out of the 1012 (10.2%), the CASAVA scores corresponding to the immune system diseases not only rank the highest, but they are also at least 10% higher than the second highest among the 24 disease categories. All these observations confirmed the relationship between MHC2TA and immune system diseases.

A recent study reported that the polymorphisms inside the IKZF1 gene are associated with systemic lupus erythematosus in the Chinese Han population (e.g., rs4917014, $p = 3 \times 10^{-6}$)[114]. In the gene body region of IKZF1, we found that the average score of immune system diseases ranks the highest among all disease categories. Also, for over 86% of variants found in the gene body or the 5 kb flanking regions of the IKZF1 gene, their CASAVA scores corresponding to the immune system disease rank either the highest or the second highest. And for 65 variants, the CASAVA scores corresponding to the immune system diseases not only rank the highest, but they are also at least 10% higher than the second highest among the 24 disease categories. In summary, we conclude that both the absolute CASAVA scores and the relative ranks among all the disease categories shed light on the level of disease risk conditioned by a given variant.

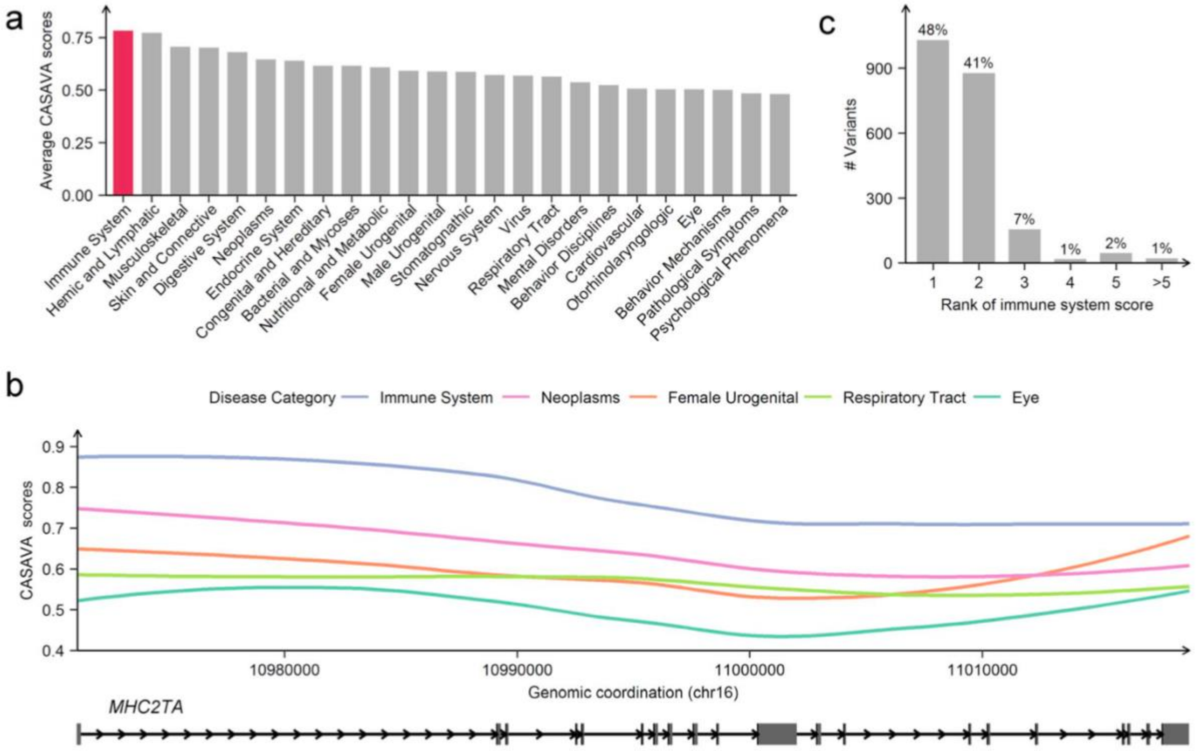


Figure 3-6 CASAVA identifies MHC2TA as an immune disease-related gene.

(a) Bar plots of average CASAVA scores inside the gene body region of MHC2TA. (b) CASAVA scores inside the gene body region of MHC2TA. The CASAVA scores are smoothed using the loess function in R. For representation, only the 1st, 6th, 11th, 16th and 21st highest CSAVA disease category scores in panel (a) are shown. (c) Numbers of variants in the gene body and flanking 5 kb regions of MHC2TA with its immune system CASAVA scores ranked in the top five among 24 categories.

3.3.12 Informative features in CASAVA

CASAVA has the potential to illuminate disease pathogenesis by ranking cell type-specific genomic or epigenomic features in terms of their relevance for predicting disease category-specific risk. Overall, we found that features related to histone marks, open chromatin and TF binding contributed more than other types of features (Figure 3-7a). This result makes sense because these features characterize the chromatin microenvironment around the variants of interest. For example, the DNase-read counts at loci containing risk variants were significantly higher than those of benign variants (Figure 3-7b, one-tailed t-test, $p < 2.2 \times 10^{-16}$). Similarly, H3K4me1 and H3K27ac counts of risk variants were significantly higher than those of benign variants (one-tailed t-test, $p < 2.2 \times 10^{-16}$ and $p = 5 \times$

10^{-15}correspondingly). In contrast, the pattern is reversed for marks of heterochromatin, such as H3K9me3 (one-tailed t-test, $p = 7 \times 10^{-8}$), suggesting that the risk variants were more likely to be found in open chromatin regions such as active enhancer and promoter regions.

We also found that the top CASAVA features often show close connections to corresponding disease categories. For example, the open chromatin features of immune-related cells such as B cell, CD4, CD8 and CD19 cells, dominate the top features in the immune disease category model (Figure 3-7c). Furthermore, risk variants associated with hemic and lymphatic diseases show depletion of open chromatin regions in blood-related cell lines such as GM12891, GM12892 and GM19239 (Figure 3-7d), indicating the tissue specificity of the hemic disease category. We also noticed that open chromatin marks—H3K4me1 and H3K27ac in the CD19 cells— are frequently selected as important features, implying that the CD19 cell type might play a key role in hemic or lymphatic traits. As shown in the literature, CD19-related therapy has been widely used to treat leukemia, which is a major disease in this hemic or lymphatic disease category[115]. Regarding bacterial infection and mycoses, we found that the closed chromatin mark H3K27me3 features in multiple cells; such cells show more depletion around risk variants than around benign variants (Figure 3-7e).

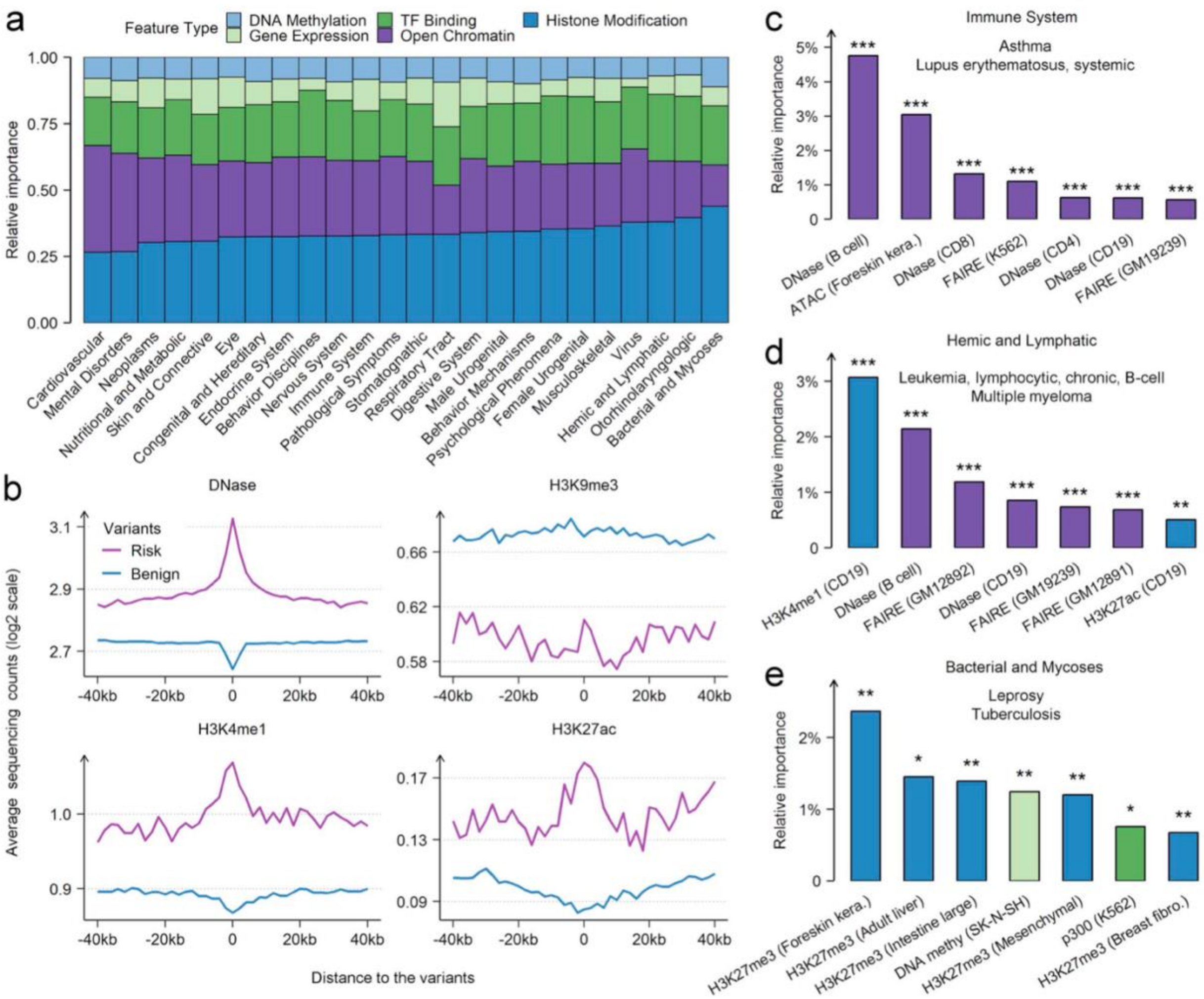


Figure 3-7 Informative features in CASAVA.

(a) Proportions of contribution by different groups of features in the 24-disease category-specific CASAVA models. (b) Selected top-ranked features of the A549 cell line for all risk and benign variants. The line plots show the average read counts (log2 scale) averaged over all risk or benign variants in the training sets. Bar charts for selected top-ranked features in models of (c) immune system diseases, (d) hemic and lymphatic diseases, (e) bacterial infection and mycoses The colors scheme is the same as the one used in panel a. P-value is calculated using the Mann–Whitney U-test between risk and benign variants. $* \, 1 \times 10^{-8} < P \leq 5 \times 10^{-2}$; $** \, 1 \times 10^{-16} < P \leq 1 \times 10^{-8}$; $*** \, P \leq 1 \times 10^{-16}$.

3.4 Discussion

In this paper, we presented CASAVA, an ensemble learning framework for disease category-specific prediction of risk variants in non-coding regions of the genome. Building on features derived from genome-wide profiling experiments, CASAVA returns risk scores for 24 disease categories for each genetic variant. Compared to existing methods, CASAVA provides more accurate prediction at the disease category level. Additionally, we found that CASAVA scores can also be used for disease-specific risk prediction; for some diseases, its performance is even better than disease-specific prediction, implying the wide-ranging applicability of CASAVA. To further improve performance for disease-specific risk prediction, we developed a transfer learning version of CASAVA, named TrCASAVA, by taking advantage of both disease specific training data as well as larger, but less specific, training data from related diseases belonging to the same disease category.

To demonstrate the utility of CASAVA, we surveyed CASAVA scores across the genome and identified genes harboring multiple variants with distinctly high CASAVA scores in a particular disease category. Among our findings, MHC2TA and IKZF1 stand out as likely to be associated with immune diseases. This connection is supported by present scientific literature. In addition to predicting scores, CASAVA also has the ability to further explore the informative features CASAVA selected during the feature selection step for each disease category. For example, a TF in a specific cell type, or a histone mark in a specific tissue type, could potentially illuminate possible disease pathogenesis or etiology.

The motivation for developing CASAVA is to find a compromise between general pathogenicity (disease-neutral) prediction and disease-specific prediction. Using a single score such as CADD score is appealing due to its simplicity, but insufficient to describe pathogenicity of diverse diseases due to their heterogeneous and complex nature. And the disease-specific approach like DIVAN is often limited by the small number of known disease-associated risk variants which is required to set up the training set. In this work, we describe CASAVA and TrCASAVA, which achieve a trade-off between generality and specificity of different diseases.

Currently, CASAVA considers 2,725 features. They belong to five broad categories: open chromatin, histone modification, TF binding, gene expression, and DNA methylation. In the future, we will include single-cell RNA-seq or ATAC-seq data as additional features as they become increasingly available, which may be used to describe the chromatin environment at the single-cell level[116, 117]. We may also use Hi-C data to capture chromatin conformation[118].

We studied the effects of different benign variants for testing. Besides TSS-matched benign variants, we also performed testing using unmatched or region-matched benign variants. In both cases, CASAVA worked better than others, followed by PAFA and GWAVA. It is worth noting that all methods perform significantly worse when tested on region-matched benign variants. The similar chromatin landscape between risk and benign variants poses more difficulties in this situation. Perhaps improving the resolution (from 200-bp bin to 100-bp or 50-bp bin) of the features will alleviate the problem.

We also explored the heterogeneity of risk variants: for example, the majority of risk variants located in the intron or intergenic regions. We first performed a separate evaluation by only using risk variants in the intron or intergenic regions. In both cases, CASAVA worked better than others. However, some methods, such as LINSIGHT or GWAVA TSS-matched score, might only be able to deal with either intron or intergenic variants. All these results pointed to the difference between the chromatin landscapes of risk variants in the intron regions and those in the intergenic regions. Hence, considering the differences between intron and intergenic regions may give better results.

There are many potential applications for CASAVA scores. Here we describe two potential applications: (1) identifying disease-associated genes and (2) exploring connections among various diseases or traits.

What is the best way to identify variants that are likely to be associated with one or more disease categories using CASAVA, especially in the case of rare variants? Or how can one identify loci that harbor risk variants for certain categories of diseases? As shown above, CASAVA provides information from different aspects. First, the higher the CASAVA score of the variant, the more elevated is the risk (presumably) associated with that disease category. In particular, the disease category with the highest score among the 24 categories is, perhaps, most worth following up on, especially because it is significantly higher than the scores from all the other categories.

CASAVA scores may also be exploited to explore relationships among different disease categories. To this end, we collected CASAVA scores for all disease categories and for all variants in chromosome 1 (as cataloged by the 1000 genomes project phase 3), which we considered as representatives of genome-wide variants; we then calculated the Pearson correlation between the vectors of scores from every pair of the 24 disease categories. We found for example, that CASAVA scores for male and female urogenital diseases are quite similar, indicating the commonalities between urogenital diseases. We also found that the bacterial infection and mycoses categories are very different from other categories. Generally, similarities between CASAVA scores of different categories are high, indicating that risk variants for different diseases indeed share some common chromatin signatures.

The overarching goal of CASAVA is to provide an alternative way to evaluate the impact of non-coding variants in terms of disease category-specific risk. Population-based approaches like GWAS, although effective and reliable for identifying disease-associated variants, their discovery power is limited by important factors such as minor allele frequencies. It has been showed that pathogenic SNVs have a wide spectrum of minor allele frequencies. For example, some SNVs are common with low penetrance, whereas other SNVs are quite rare but show high penetrance. SNVs in the later categories may not be identified by GWAS even with all the populations in the world.

CASAVA aims to provide an alternative approach to predict and potentially identify SNVs that associated with various categories of human diseases and traits that traditional GWASs are unable to achieve. CASAVA is able to accomplish this by bring in molecular features that are not used in classical GWAS. We think this is important because all human diseases develop with certain biological mechanisms. Such information can be found in molecular level, genome-wide profiling assays such as ChIP-seq, ATAC-seq data.

In personal sequencing studies, we may discover an ultra-rare SNV that has never been implicated by any genetic study. However, from its local genomic and epigenomic profiles, we may be able to predict that it is capable of conveying significant risk to one or more disease categories. We believe such information can be important in translational research. And the information obtained from CASAVA is complimentary to what GWAS can provide us.

We acknowledge that the accuracy and specificity of CASAVA still have much room for improvement at the moment. But we believe that our results showed that the strategy works in principal and performs better than other competitors. In many machine learning applications, the quality of the training data and features play important roles in its performance.

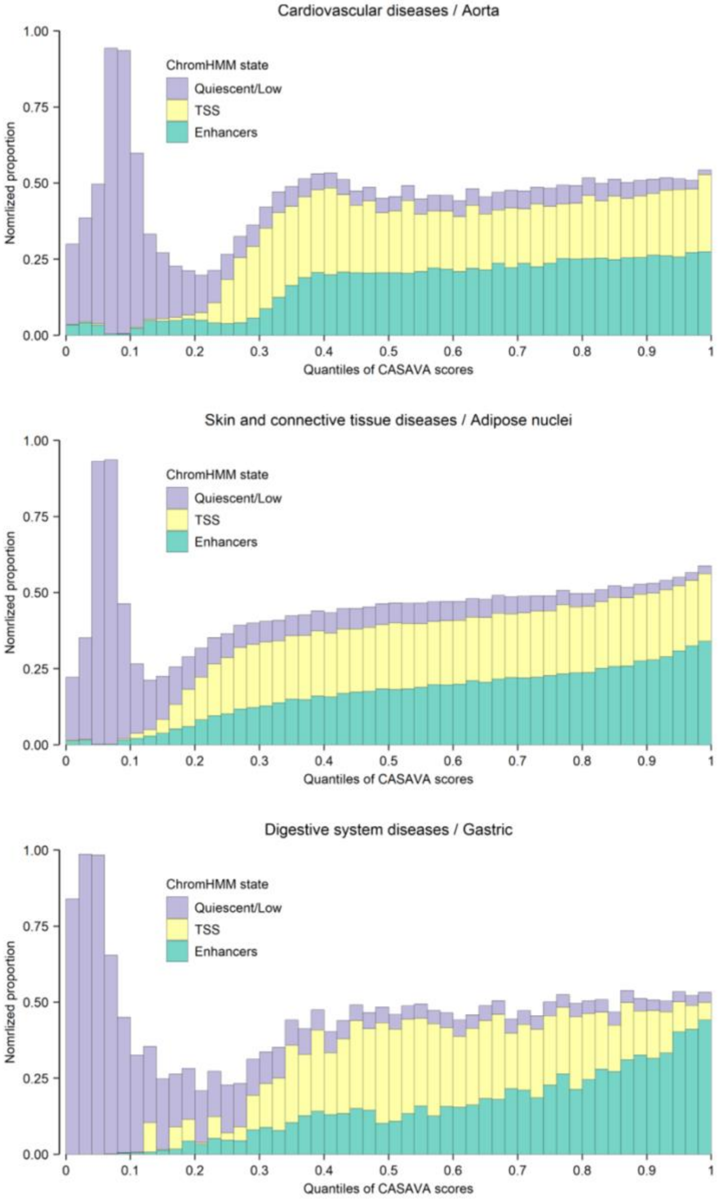
Training data used in CASAVA are collected from PheGenI, despite complications such as not all GWAS SNVs are reproducible and the top-ranked index SNP in a locus may not be the causal one, we believe the proportion of bona fide disease-associated variants is much higher than that in the control set. To make the training set more reliable, we also change the p-value threshold from 10^{-3} (used by the PheGenI database) to 10^{-4}.

As of features, it is important to recognize that new and high-quality data are being continuously generated and made publicly available. With the fast-evolving technologies like single cell technologies. More diverse and informative features will become available, and they will help improve the performance of CASAVA.

An interesting question is whether CASAVA can help with fine mapping. Due to the limited resolution of most features used, and the fact that training sets are based on GWAS results which are limited by linkage disequilibrium (LD), CASAVA is not suitable to do fine mapping. However, since typically LD extends much longer than genomic or epigenomic signals (limited by the experimental assays, such as the fragment size), hence using CASAVA scores, we should be able to narrow the association locus to a genomic interval much smaller than the LD block containing the GWAS variants.

CASAVA score is assigned to the locus of the genetic variant at a 200-bp resolution, not the variant per se. CASAVA assigns a risk score for every 200-bp bin in the genome, using the local genomic and epigenomic profiles of the position. There are multiple existing methods to segment and annotate the genome[119-124], yet CASAVA is the first to provide disease category-specific risk prediction. An interesting question is whether there is any connection between CASAVA scores and these annotations. To explore we tested enrichment of various chromatin states of relevant tissues in selected disease categories and indeed, we found significant enrichment of enhancers and TSS proximal chromatin states (Figure 3-8).

Additionally, one could also test for enrichment of disease-related transcription factor binding motifs or genes belong to disease-related pathways or gene sets using existing tools[125-127].


Cardiovascular diseases / Aorta

Skin and connective tissue diseases / Adipose nuclei

Digestive system diseases / Gastric

Figure 3-8 Clear trend of increasing enrichment of regulatory chromatin states (TSS and enhancers) in the aorta tissue type with higher CASAVA scores of "cardiovascular diseases, especially for enhancers.

In Figure 3-7a, we notice that variation in the contribution of the five feature types among the 24 disease categories. For example, it seems that gene expression plays an important role in respiratory tract whereas open chromatin is less important than histone modification for bacterial and mycoses. The order of overall importance among the five categories of features is as follows: histone modification, open chromatin, TF binding, gene expression and DNA methylation. The overall pattern can be partially explained by the fact that the number of features is following roughly the same order. Additionally, two feature types: DNA methylation and TF binding are relatively stable, and the three other feature types vary substantially. An interesting phenomenon is that the gene expression features, and open chromatin features sum up roughly the same. We hypothesize that the gene expression features are important for SNVs near gene, and open chromatin features are important for SNVs farther away from genes.

Admittedly, there will be information loss in the process of assigning diseases to disease categories due to our incomplete understanding of the disease processes. Despite this, we believe it is beneficial to use disease category-level annotation. This is because, first, there are too many diseases, it is cumbersome to annotate risk for every single disease. Second, annotation based on ML strategy is not possible for most diseases because there is insufficient training data. Adopting disease category annotation, a vector of 24 scores is sufficient. And at the disease category-level, there are much more training data available for each category. For future work, we will work on fine-tuning disease category definition. A useful resource is disease ontology (DO)[128]. We will also explore how to combine related diseases or disease categories based on DO for reasonable and sufficient data utilization[129].

To make CASAVA more accessible and easier to use, we built a web server (http://zhanglabtools.org/CASAVA), along with visualization tools, for retrieving CASAVA scores. Additionally, we provided pre-computed whole-genome CASAVA scores and an easy-to-use R script for scoring a large number of variants.

In summary, this study presents a novel ensemble learning framework, CASAVA, for predicting disease category-specific risk variants in non-coding regions of the genome. Compared to ten different scoring methods, CASAVA demonstrates the best overall results in terms of both disease category-specific and disease-specific prediction. Additionally, better results can be achieved when additional known risk variants from related diseases are added under a transfer learning framework. The new algorithm, TrCASAVA, further demonstrates the advantage of pooling together risk variants from similar diseases to boost the performance. Using MHC2TA and IKZF1 genes as examples, CASAVA shows the potential of identifying novel disease- associated variants or genes. In order to make CASAVA easily accessible, we built a web portal to allow easy browsing and querying of CASAVA scores (http://zhanglabtools.org/CASAVA).

Chapter 4 *DRAFT*: Disease Risk Annotation with Few shoTs learning

4.1 Introduction

Few shots learning (FSL) models have been successfully utilized in various computer vision applications, including signature verification[130] and face recognition[131]. More recently, FSL applications have emerged as competitive machine learning tools in broader fields of studies; however, the use of FSL in life sciences is still rare. Here, we propose to apply Siamese neural networks, an FSL technique, to learn a distance function which is capable of mapping risk variant of interest into low dimensional embedded space, such that a pair of risk variants are closer to each other in the embedded space than a risk and a benign variant. We implemented the strategy into a computational tool named DRAFT (Disease Risk Annotation with Few shoTs learning), an end-to-end package applicable to both genetic and epigenetic variant annotation. Test on real-data demonstrated superior performance over deep-learning-based classification approaches.

Identifying disease-susceptible variants is important for delineating the mechanism of complex diseases. Over the last two decades, thousands of genome-wide association sties(GWASs)[132] and epigenome-wide association studies(EWASs)[133] have identified tens of thousands of genetic and epigenetic variants that are associated with hundreds of complex diseases. Despite extensive findings, more remains to be discovered. A key to recognize more such variants is to understand what makes a variant to be disease-associated. This can be framed as a supervised learning problem. Indeed, many machine learning approaches have already been developed to predict the pathogenicity of genetic variants and epigenetic modifications[9, 12, 59, 91, 95, 98, 134]. Existing variant annotation methods dichotomize all variants into pathogenic or benign. A caveat of such an over-simplified strategy is that quantitative information associated with variants, such as p-value or effect size, is being overlooked. In this study, we propose an alternative strategy for variant annotation by reframing the problem as a distance metric learning problem[135], such that quantitative information such as p-values or effect sizes can be utilized to better characterize the disease risk of these variants. Achieving top performance using sophisticated machine learning algorithms such as deep learning requires a large number of high-quality training samples. This is not possible for disease-specific risk evaluation[12, 136] for most diseases. A key advantage of metric learning is that these methods can return decent performance with limited training data[137], which is why these methods are also referred to as few-shot learning methods. Supervised learning with limited training data is important for many bioinformatics applications because experimentally validated training data is often scarce.

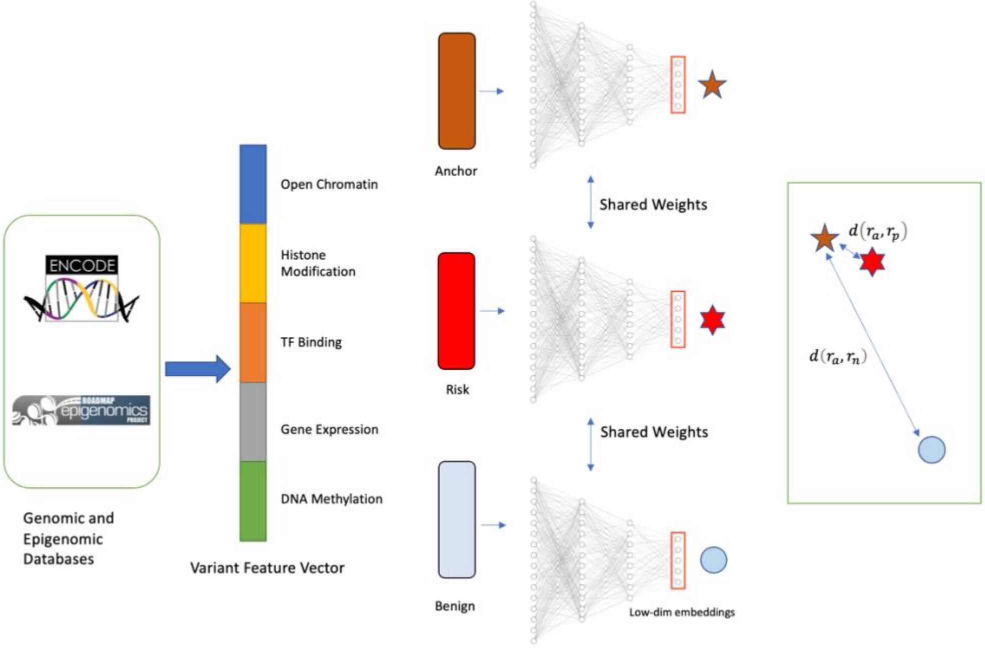


Figure 4-1 Workflow of DRAFT with triplet loss.

4.2 Methods

4.2.1 Data collection and preprocessing

Like existing variant annotation methods, we use genomic and epigenomic profiles as features. To be specific, we first downloaded mapped reads from the ENCODE[2] and the REMC[1] project. Read counts from biological replicates were merged into a single feature. In total, we assembled 2,496 genome-wide features, which can be roughly divided into five groups: open chromatin, histone modification, TF binding, gene expression and DNA methylation. The detailed number of features can be found in Table 4-1.

To simplify the calculation, we divided the entire genome into 200-bp bins and calculated the normalized mapped read counts for each bin. We stored the resulting features in a 15,685,849 by 2,496 matrix. For this matrix, each row represents a 200-bp bin, and each column represents a genomics or epigenomics feature. For a biomarker of interest, we first find which bin the biomarker falls into, then retrieve the corresponding feature vectors.

Assay Type	Number of features
Histone ChIP-seq	1,002
Transcription Factor (TF) ChIP-seq	571
RNA polymerase binding ChIP-seq	49
DNase-seq	153
FAIRE-seq	31
Total RNA-seq	243
ATAC-seq	66
WGBS	381
Total	2,496

Table 4-1 Number of features of each experiment assay.

We test the metric learning strategy on both genetic variants and epigenetic variants. For genetic variants, we collected GWAS-identified risk variants of 89 diseases from PheGenI[88] datasets spanning a wide range of complex diseases and traits. The same set of diseases are studied in Cao et al.[138] . For control sets consist of benign variants, we used a similar strategy as in the GWAVA-TSS method[9]. We first downloaded all non-coding variants in the 1000 Genomes Project[89] phase 1 release. To minimize the chance that a benign variant would be disease-implicated, we excluded all variants found within 1kb of any of variants found in the PheGenI[88] database. Next, we excluded variants with minor allele frequency (MAF) less than 1% to match the allele frequency range of the risk variants. Finally, we sampled ten times more benign variants than risk variants and required the benign variants to have roughly the same distances to the nearest transcription start sites (TSS) with risk variants.

For epigenetic variants, following Huang et al.[136], we used the raw beta values and phenotype of information of 717 patients from the ROS/MAP cohort[41] from De Jager et al.[28] and then performed EWAS on six Alzheimer's Disease (AD) related traits, including: beta-amyloid density, Braak staging, the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) score, cognitive trajectory, global AD pathology, and neurofibrillary tangle density. To construct the control set, we firstly selected CpGs with significance level of greater than 0.4 from differential methylation test. We then selected ten benign CpGs that have most similar methylation levels for each risk CpG.

4.2.2 Triplet Loss and Lifted Structured Loss

Triplet loss was originally proposed in the FaceNet[131] paper. Each training triplet is formed by an anchor variant, a positive(risk) variant and a negative(benign) variant. The training objective is to maximize the distance between the anchor and the negative variant $\left(d(r_a, r_n)\right)$ by at least m units more than the distance between the anchor and the positive variant $(d(r_a, r_p))$. Mathematically, the triplet loss can be calculated as follows:

$$L(r_a, r_p, r_n) = \max\left(0, m + d(r_a, r_p) - d(r_a, r_n)\right)$$

r_a, r_p, r_n denotes the low-dimensional representation of the anchor, the risk and benign biomarker, m is the margin. It is crucial to select benign samples that are hard to distinguish from risk variants, otherwise no gradients would be generated for backpropagation. During training, hard triplets and semi-hard are chosen for each training batch. Hard triplets contain negative biomarker closer to the anchor than the positive and semi-hard contains negative further from the anchor than the positive, but the difference of distance is less than the margin m.

Lifted structured loss[139] further utilizes all the pairwise edges within one training batch. The dense pairwise distance matrix is computed for every training batch. Essentially, for each risk pair, this type of loss function involves the distance between the risk sample and every other negative sample, the smoothed version of lifted structured loss is as follows:

$$L_{struct} = \frac{1}{2|P|} \sum_{(i,j)\in P} \max\left(0, L_{struct}^{(ij)}\right)^2$$

$$where\ L_{struct}^{(ij)} = D_{ij} + \log\left(\sum_{(i,k)\in N} \exp(m - D_{ik}) + \sum_{(j,l)\in N} \exp\left(m - D_{jl}\right)\right)$$

m is margin parameter to push the distance of risk-benign pair further. D_{ij} denotes the distance between risk sample i and j, D_{ik} denotes the distance between risk sample i and benign sample k. P contains the set of positive pairs and N is the set of negative pairs.

4.2.3 Implementation details

We used the PyTorch[140] package for training and testing multi-layer perceptron (MLP) and ResNet[141] with triplets, lifted structured loss and classification with cross entropy loss settings. Maximum training epochs was set at 500. Early stopping with the maximum tolerance of 10 epochs was applied to prevent over- training. Learning rate scheduler was set to reduce the learning rate by the factor of 0.1 when the validation AUC stops increasing for one epoch. The batch size was set to 128 for triplet and classification with cross entropy loss and 32 for lifted structured loss. For triplet and lifted structured loss, we use mining strategy to select all hard negatives and semi-hard negatives for calculating the loss. Random seed was set to ensure the training and validation set split was reproducible and same for model comparison.

4.3 Results

4.3.1 Evaluation and performance comparison

Our preliminary experiments compare the performance of three models constructed using Siamese neural networks of a) ResNet, b) multi-layer perceptron (MLP) with triplet loss and c) fully connected neural network trained with classification loss (cross-entropy loss). Model c is the baseline model. In order to compare the performance of few shots learning models (a and b) and the traditional classification model (c), after the training is completed, for the i^{th} unseen testing sample x_{test_i}, we calculate the distance between x_{test_i} and the positive sample groups (assume the number of positive samples in the training set is k) in the training set and denotes the distance set as $\left\{d_{i_{pos}} \middle| d\left(x_{test_i}, x_{train_{pos_1}}\right), \left(x_{test_i}, x_{train_{pos_2}}\right), \dots, \left(x_{test_i}, x_{train_{pos_k}}\right)\right\}$, as well the distance between x_{test_i} and the negative sample groups (assume number of negative samples in the training set is j) in the training set and denotes the distance set as $\left\{d_{i_{neg}} \middle| d\left(x_{test_i}, x_{train_{neg_1}}\right), \left(x_{test_i}, x_{train_{neg_2}}\right), \dots, \left(x_{test_i}, x_{train_{neg_j}}\right)\right\}$. Then, the ratio of the mean of set $d_{i_{neg}}$ and $d_{i_{pos}}$ is calculated as the prediction. A larger ratio suggests the unseen variant of interest is closer to positive samples in the training set and thus should be predicted as positive (risk variant). In contrast, a smaller ratio suggests the unseen variant is more likely to be benign. The AUC calculated based on the ratio predictions is used to compare with the baseline binary classification model.

We performed 5-fold cross validation on our GWAS and EWAS dataset. For GWAS part, we compared our result with eleven existing genomic annotation tools. From a total of 14 methods, Siamese network-based methods performed best in 55 out of 89 diseases if applying triplet loss, and 58 out of 89 diseases if applying structured lifted loss. The average cv-AUC for Siamese-based model with triplet loss and structured lifted loss are similar (Table 4-2(GWAS dataset) Mean cv-AUC of 14 methods across 89 diseases dataset.). In that case, triplet loss is preferred due to its advantage in computation efficiency. The following experiments comparison for the Siamese network-based models are all based on triplet loss. The performance of Siamese network-based ResNet is significantly better than the baseline MLP classification model (mean cv-AUC: 0.657 vs 0.615; $p = 2.665 \times 10^{-7}$, one-side Wilcoxan rank sum test) and the same conclusion holds for the comparison of Siamese network-based MLP and baseline MLP classification model (mean cv-AUC: 0.658 vs 0.615; $p = 2.810 \times 10^{-7}$, one-side Wilcoxan rank sum test). The detailed performance comparison for all 14 methods can be found in Table 4-2 and Figure 4-2. The detailed performance of 14 methods for each disease can be found in Figure 4-3. Due to the fact that there are only a few annotation tools existing for the EWAS studies, we only compared model performance to the baseline model. The performance of Siamese network-based ResNet is better than the baseline MLP classification model (mean cv-AUC: 0.824 vs 0.811) and the same conclusion holds for the comparison of Siamese network-based MLP and baseline MLP classification model (mean cv-AUC: 0.827 vs 0.811). However, we do not observe statistically significant difference in these two comparisons. More future experiments are needed to illustrate the usefulness of Siamese-based methods in the application of EWAS datasets. The detailed performance comparison between Siamese-based methods and baseline model can be found in Table 4-3.

In addition, we also performed experiment with subsampling to see if Siamese network-based methods perform better when the number of training samples is limited. We performed the experiment on Myocardial Infarction GWAS dataset, one of the datasets with relatively large number of risk training samples (N=698). We can observe from Table 4 that although all three methods suffer worsen performance when the sampling rate increases, the performance margin between Siamese network-based methods and MLP classification baseline model become larger when only extreme small fraction of risk samples are available.

Method	AUC
Siamese-ResNet (Triplet Loss)	0.657
Siamese-MLP (Triplet Loss)	0.658
Siamese-ResNet (Lifted Structured Loss)	0.652
Siamese-MLP (Lifted Structured Loss)	0.658
MLP-classification	0.615
PAFA	0.607
GWAVA-TSS	0.598
GWAVA-Region	0.584
GWAVA-Unmatched	0.530
Eigen-PC	0.572
Eigen	0.568
GenoCanyon	0.568
CADD	0.560
LINSIGHT	0.556
FATHMM-MKL	0.527
DANN	0.519

Table 4-2(GWAS dataset) Mean cv-AUC of 14 methods across 89 diseases dataset.

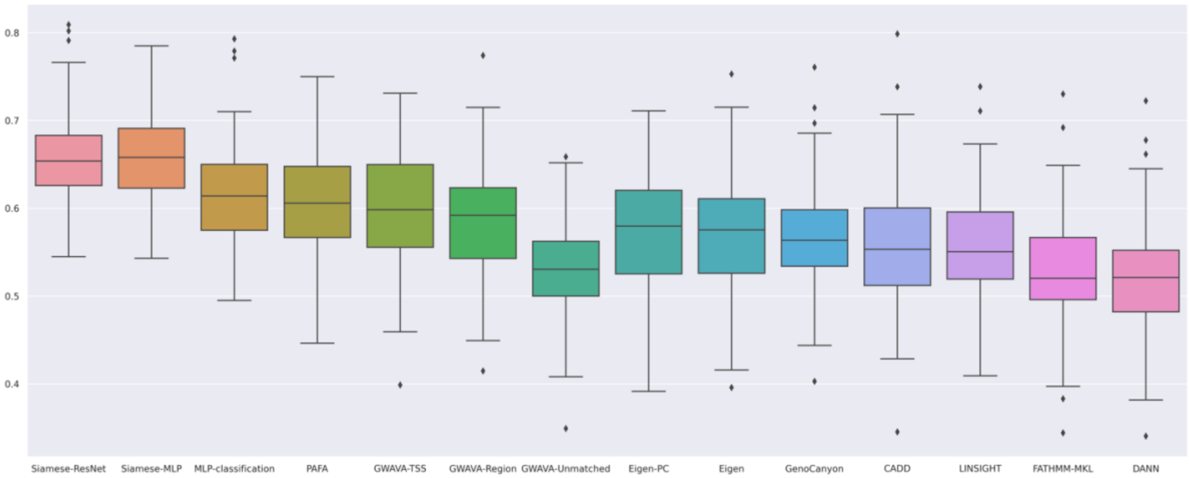


Figure 4-2 Performance comparison box plot of Siamese based methods, baseline classification-based method and 12 other existing genomic annotation tools across 89 diseases for GWAS dataset

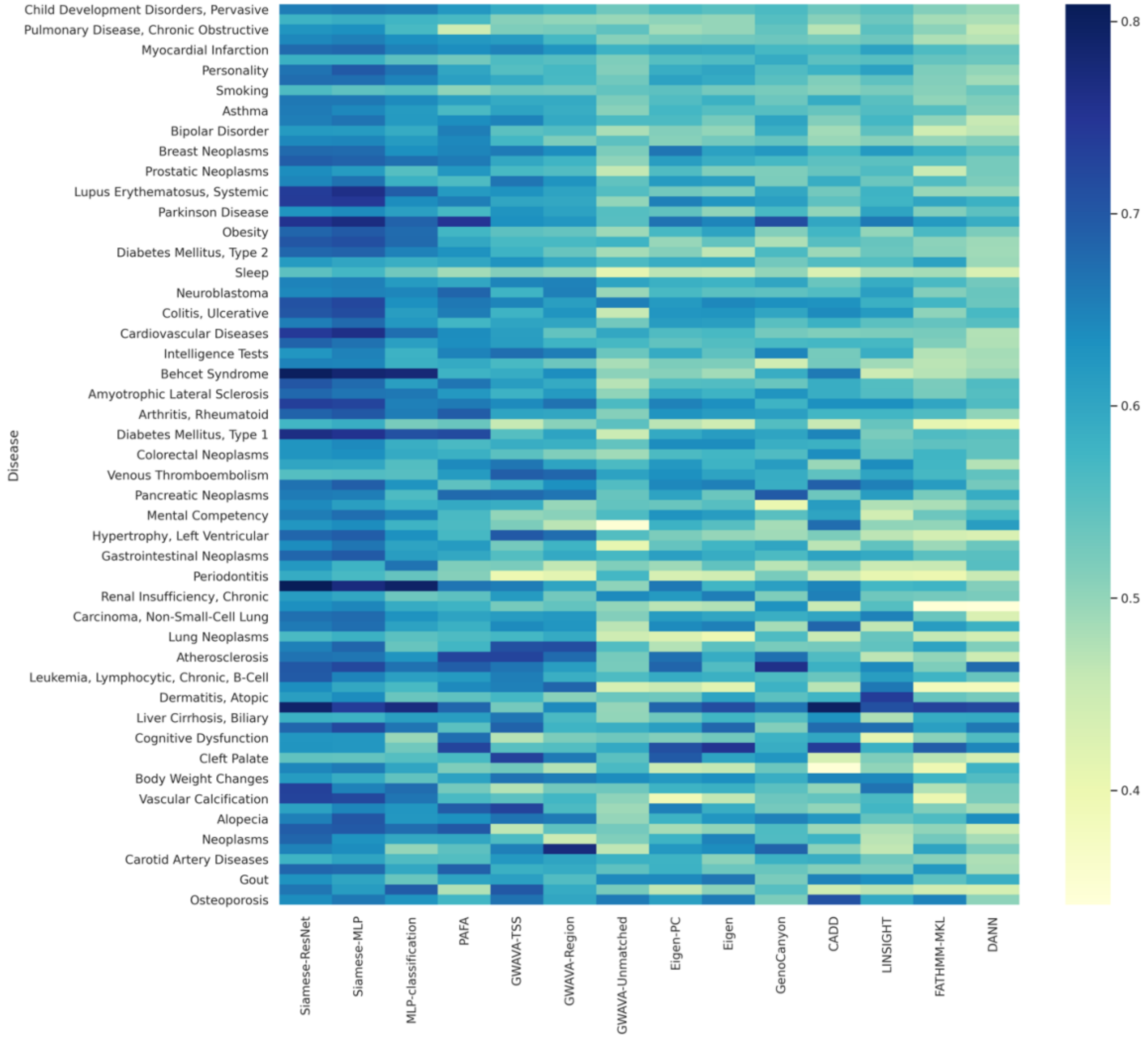


Figure 4-3 cv-AUC heatmap of 14 methods for each of 89 diseases.

Trait	Siamese-ResNet	Siamese-MLP	MLP-classification
Beta-amyloid	0.762	0.762	0.749
Braak staging scores	0.802	0.803	0.782
CERAD	0.796	0.801	0.776
Cognitive decline trajectory	0.793	0.797	0.771
Global Pathology	0.844	0.853	0.841
Neurofibrillary tangles	0.946	0.948	0.949

Table 4-3 (EWAS dataset) Mean cv-AUC of Siamese-based methods and the baseline method for 6 AD related EWAS datasets.

4.3.2 Conclusion

In summary, in this work, we explore applying few shots learning on disease risk annotation for GWAS and EWAS studies. Our experiments show that few shorts learning based risk annotation models have better performance compared to many existing tools and the baseline classification model. The promising results shows that few shots learning method can be a trustworthy alternative for the traditional classification framework in this research field, especially when the number of training samples is limited. In addition, deep networks can effectively exploit the relationships between omics data without performing feature selection and thus provides a neat end-to-end solution. The omics genome-wide features we preprocessed from the ENCODE and the ROADMAP project can also be a valuable resource for other genomic/epigenomic studies that would like to utilize omics data.

Chapter 5 Future Works

The continuous evolving sequencing technologies have enabled endless new possibility for more precise and comprehensive genome annotation. The single-cell sequencing technology, which was named as the Method of the Year in 2013[142], and the single-cell multimodal omics sequencing technology, which was named as the Method of the Year in 2019[143], have enabled us to explore biological insight at the cellular level. As one of its applications, single-cell sequencing is capable of revealing somatic mutations and structural changes in cancer cells, which tend to have high mutation rates, allowing researchers to identify biomarkers for the prediction and validation of small molecules that target drug-tolerant cells[144]. Incorporating single-cell omics data (e.g., single-cell RNA seq, single-cell ATAC seq, single cell DNA methylome seq, etc.) into genome annotation can give us deeper insight from a different perspective than traditional bulk sequencing data and provide greater transparency in discovering informative features, thus allowing us to demonstrate further the relevance between our annotation results and biological discoveries to researchers who are interested in the interpretation behind annotation scores.

Besides the current sequencing data involved in the modeling, cohorts such as UK Biobank[145] have provided a large volume of individual level data including electronic health records (EHR) and other individual phenotypes that can be combined with the genome annotation results for personalized risk prediction for certain diseases. In addition to the individual phenotypes data, as most diseases are both genetically and environmentally influenced[146], including environmental exposure data can complement the sequencing data we current have. From the modeling perspective, it is reasonable to design a model structure with two independent blocks that process the genomic/epigenomic profiles and individual level phenotypes data separately. Essentially, these two blocks are designed as feature extractor modules to handle these two types of features separately, and the extracted high-level features can then be combined in the final output module to make the final prediction. It is worthwhile to note that the usage of individual level data should adhere to strict data privacy regulation and subject to approval from the Institutional Review Boards (IRBs).

In addition to the data from additional types of sequencing assays, newly emerged advanced machine learning algorithms also enable the analysis of more types of input data. Modern deep learning models that can handle sequence-like data would be a promising direction for the future genome annotation tasks. The current proposed approaches in my thesis assign the features to the biomarker of interest based solely on its genomic coordinate. However, there is a spatial interaction and connection between genomic activities, and the activities that take place in a given genomic region can be influenced by those that occur in regions downstream or upstream. For example, SNPs found in gene bodies are not the only ones that affect gene transcription, SNPs located in intergenic regulatory sequences, such as enhancers, may also interfere with normal activities of nearby regulatory elements and thereby impact gene expression[147]. Hence, it is reasonable to include the nearby genomic and epigenomic profiles of the variants into the modeling process. The development of Recurrent neural network (RNN) with Long Short-Term Memory (LSTM)[148] and Gated Recurrent Units (GRU)[149], as well as an even more powerful self-attention based model Transformer[150] has revolutionized natural language processing and computer vision. These techniques are capable of training with long sequences data without experiencing gradient exploding or gradient vanishing. Moreover, these models can also be extended directly to include the DNA sequence of four types of nucleotides. RNN and Transformer are able to harness information from the DNA sequence patterns and generate high-level features that can be further utilized for the final prediction of the annotation outcome. As the next step, we would like to explore taking advantage of these new methods to update EAWASplus, CASAVA and DRFAT to further improve their performance.

Bibliography

1. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR. The NIH roadmap epigenomics mapping consortium. Nature biotechnology. 2010;28(10):1045-8.

2. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

3. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome research. 2007;17(6):877-85.

4. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols. 2010;2010(2):pdb. prot5384.

5. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature reviews genetics. 2009;10(10):669-80.

6. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics. 2009;10(1):57-63.

7. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC‐seq: a method for assaying chromatin accessibility genome‐wide. Current protocols in molecular biology. 2015;109(1):21.9. 1-.9. 9.

8. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94. Epub 2018/10/30. doi: 10.1093/nar/gky1016. PubMed PMID: 30371827; PMCID: PMC6323892.

9. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3):294-6. Epub 2014/02/04. doi: 10.1038/nmeth.2832. PubMed PMID: 24487584; PMCID: PMC5015703.

10. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.

11. Chen J, Schwarz E. BioMM: Biologically-informed Multi-stage Machine learning for identification of epigenetic fingerprints. arXiv preprint arXiv:171200336. 2017.

12. Chen L, Jin P, Qin ZS. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. Genome Biol. 2016;17(1):252. Epub 2016/12/08. doi: 10.1186/s13059-016-1112-z. PubMed PMID: 27923386; PMCID: PMC5139035.

13. Hoyert DL, Heron MP, Murphy BS, Kung H. National Vital Statistics Reports - Deaths: Final Data for 2003. Center for Disease Control and Prevention: U.S. Department of Health and Human Services; 2006. p. 1-16.

14. Wingo TS, Lah JJ, Levey AI, Cutler DJ. Autosomal recessive causes likely in early-onset Alzheimer disease. Archives of neurology. 2012;69(1):59-64. Epub 2011/09/14. doi: 10.1001/archneurol.2011.221. PubMed PMID: 21911656; PMCID: PMC3332307.

15. Gatz M, Pedersen NL, Berg S, Johansson B, Johansson K, Mortimer JA, Posner SF, Viitanen M, Winblad B, Ahlbom A. Heritability for Alzheimer's disease: the study of dementia in Swedish twins. J Gerontol A Biol Sci Med Sci. 1997;52(2):M117-25. Epub 1997/03/01. PubMed PMID: 9060980.

16. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010-22. Epub 2011/05/18. doi: 10.1101/gad.2037511. PubMed PMID: 21576262; PMCID: PMC3093116.

17. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6(8):597-610. Epub 2005/09/02. doi: 10.1038/nrg1655. PubMed PMID: 16136652.

18. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102(30):10604-9. Epub 2005/07/13. doi: 10.1073/pnas.0500398102. PubMed PMID: 16009939; PMCID: PMC1174919.

19. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440-6. Epub 2010/03/12. doi: 10.1101/gr.103606.109. PubMed PMID: 20219944; PMCID: PMC2847747.

20. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20(4):434-9. Epub 2010/03/12. doi: 10.1101/gr.103101.109. PubMed PMID: 20219945; PMCID: PMC2847746.

21. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. Nature. 2010;465(7299):721-7. Epub 2010/06/11. doi: 10.1038/nature09230. PubMed PMID: 20535201.

22. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. 2006;7(1):21-33. Epub 2005/12/22. doi: 10.1038/nrg1748. PubMed PMID: 16369569.

23. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115. Epub 2013/10/22. doi: 10.1186/gb-2013-14-10-r115. PubMed PMID: 24138928; PMCID: PMC4015143.

24. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, Shchetynsky K, Scheynius A, Kere J, Alfredsson L, Klareskog L, Ekstrom TJ, Feinberg AP. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nature biotechnology. 2013;31(2):142-7. doi: 10.1038/nbt.2487. PubMed PMID: 23334450.

25. Edris A, den Dekker HT, Melen E, Lahousse L. Epigenome-wide association studies in asthma: A systematic review. Clin Exp Allergy. 2019. Epub 2019/04/23. doi: 10.1111/cea.13403. PubMed PMID: 31009112.

26. Everson TM, Punshon T, Jackson BP, Hao K, Lambertini L, Chen J, Karagas MR, Marsit CJ. Cadmium-Associated Differential Methylation throughout the Placental Genome: Epigenome-Wide Association Study of Two U.S. Birth Cohorts. Environ Health Perspect. 2018;126(1):017010. Epub 2018/01/27. doi: 10.1289/EHP2192. PubMed PMID: 29373860.

27. Kingsley SL, Eliot MN, Whitsel EA, Huang YT, Kelsey KT, Marsit CJ, Wellenius GA. Maternal residential proximity to major roadways, birth weight, and placental DNA methylation. Environ Int. 2016;92-93:43-9. Epub 2016/04/09. doi: 10.1016/j.envint.2016.03.020. PubMed PMID: 27058926; PMCID: PMC4913202.

28. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, Eaton ML, Keenan BT, Ernst J, McCabe C, Tang A, Raj T, Replogle J, Brodeur W, Gabriel S, Chai HS, Younkin C, Younkin SG, Zou F, Szyf M, Epstein CB, Schneider JA, Bernstein BE, Meissner A, Ertekin-Taner N, Chibnik LB, Kellis M, Mill J, Bennett DA. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17(9):1156-63. Epub 2014/08/19. doi: 10.1038/nn.3786. PubMed PMID: 25129075; PMCID: PMC4292795.

29. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, Troakes C, Al-Sarraj S, Burrage J, Macdonald R, Condliffe D, Harries LW, Katsel P, Haroutunian V, Kaminsky Z, Joachim C, Powell J, Lovestone S, Bennett DA, Schalkwyk LC, Mill J. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat Neurosci. 2014;17(9):1164-70. doi: 10.1038/nn.3782. PubMed PMID: 25129077; PMCID: PMC4410018.

30. Do C, Lang CF, Lin J, Darbary H, Krupska I, Gaba A, Petukhova L, Vonsattel JP, Gallagher MP, Goland RS, Clynes RA, Dwork A, Kral JG, Monk C, Christiano AM, Tycko B. Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. Am J Hum Genet. 2016;98(5):934-55. doi: 10.1016/j.ajhg.2016.03.027. PubMed PMID: 27153397; PMCID: PMC4863666.

31. Watson CT, Roussos P, Garg P, Ho DJ, Azam N, Katsel PL, Haroutunian V, Sharp AJ. Genome-wide DNA methylation profiling in the superior temporal gyrus reveals epigenetic signatures associated with Alzheimer's disease. Genome Med. 2016;8(1):5. doi: 10.1186/s13073-015-0258-8. PubMed PMID: 26803900; PMCID: PMC4719699.

32. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. Genomics data. 2016;9:22-4.

33. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16:14. Epub 2015/01/27. doi: 10.1186/s13059-015-0581-9. PubMed PMID: 25616342; PMCID: PMC4389802.

34. De Jager PL, Shulman JM, Chibnik LB, Keenan BT, Raj T, Wilson RS, Yu L, Leurgans SE, Tran D, Aubin C, Anderson CD, Biffi A, Corneveaux JJ, Huentelman MJ, Alzheimer's Disease Neuroimaging I, Rosand J, Daly MJ, Myers AJ, Reiman EM, Bennett DA, Evans DA. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol Aging. 2012;33(5):1017 e1-15. Epub 2011/11/08. doi: 10.1016/j.neurobiolaging.2011.09.033. PubMed PMID: 22054870; PMCID: PMC3307898.

35. Smith RG, Hannon E, De Jager PL, Chibnik L, Lott SJ, Condliffe D, Smith AR, Haroutunian V, Troakes C, Al‐Sarraj S. Elevated DNA methylation across a 48‐kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimer's & Dementia. 2018;14(12):1580-8.

36. Beach TG, Adler CH, Sue LI, Serrano G, Shill HA, Walker DG, Lue L, Roher AE, Dugger BN, Maarouf C. Arizona study of aging and neurodegenerative disorders and brain and body donation program. Neuropathology: official journal of the Japanese Society of Neuropathology. 2015;35(4):354.

37. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, Klein H-U, White CC, Peters MA, Lodgson B. A multi-omic atlas of the human frontal cortex for aging and Alzheimer’s disease research. Scientific data. 2018;5:180142.

38. Li T, Kim A, Rosenbluh J, Horn H, Greenfeld L, An D, Zimmer A, Liberzon A, Bistline J, Natoli T. GeNets: a unified web platform for network-based genomic analyses. Nature methods. 2018;15(7):543-6.

39. Wang JZ, Xia YY, Grundke-Iqbal I, Iqbal K. Abnormal hyperphosphorylation of tau: sites, regulation, and molecular mechanism of neurofibrillary degeneration. J Alzheimers Dis. 2013;33 Suppl 1:S123-39. Epub 2012/06/20. doi: 10.3233/JAD-2012-129031. PubMed PMID: 22710920.

40. Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res. 2012;9(6):646-63. Epub 2012/04/05. PubMed PMID: 22471867; PMCID: PMC3439198.

41. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis. 2018;64(s1):S161-S89. Epub 2018/06/06. doi: 10.3233/JAD-179939. PubMed PMID: 29865057; PMCID: PMC6380522.

42. Smith R, Pishva E, Shireby G, Smith AR, Roubroeks JA, Hannon E, Wheildon G, Mastroeni D, Gasparoni G, Riemenschneider M. Meta-analysis of epigenome-wide association studies in Alzheime's disease highlights 220 differentially methylated loci across cortex. BioRxiv. 2020.

43. Brokaw DL, Piras IS, Mastroeni D, Weisenberger DJ, Nolz J, Delvaux E, Serrano GE, Beach TG, Huentelman MJ, Coleman PD. Cell death and survival pathways in Alzheimer's disease: an integrative hypothesis testing approach utilizing-omic data sets. Neurobiology of Aging. 2020;95:15-25.

44. Bharadwaj P, Martins RN. PRKAG2 Gene Expression Is Elevated and its Protein Levels Are Associated with Increased Amyloid-β Accumulation in the Alzheimer's Disease Brain. Journal of Alzheimer's Disease. 2020;74(2):441-8.

45. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9. Epub 2013/01/15. doi: 10.4161/epi.23470. PubMed PMID: 23314698; PMCID: PMC3592906.

46. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP, Ressler KJ, Conneely KN. Accounting for population stratification in DNA methylation studies. Genet Epidemiol. 2014;38(3):231-41. Epub 2014/01/31. doi: 10.1002/gepi.21789. PubMed PMID: 24478250; PMCID: PMC4090102.

47. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3. Epub 2012/01/20. doi: 10.1093/bioinformatics/bts034. PubMed PMID: 22257669; PMCID: PMC3307112.

48. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics. 2012;28(9):1280-1. Epub 2012/03/28. doi: 10.1093/bioinformatics/bts124. PubMed PMID: 22451269; PMCID: PMC3577110.

49. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. Epigenetics. 2013;8(3):290-302. Epub 2013/02/22. doi: 10.4161/epi.23924. PubMed PMID: 23426267; PMCID: PMC3669121.

50. Bennett DA, Schneider JA, Arvanitakis Z, Kelly JF, Aggarwal NT, Shah RC, Wilson RS. Neuropathology of older persons without cognitive impairment from two community-based studies. Neurology. 2006;66(12):1837-44. Epub 2006/06/28. doi: 10.1212/01.wnl.0000219668.47116.e6. PubMed PMID: 16801647.

51. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82(4):239-59. Epub 1991/01/01. doi: 10.1007/bf00308809. PubMed PMID: 1759558.

52. Mirra SS, Heyman A, McKeel D, Sumi SM, Crain BJ, Brownlee LM, Vogel FS, Hughes JP, van Belle G, Berg L. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. Neurology. 1991;41(4):479-86. Epub 1991/04/01. doi: 10.1212/wnl.41.4.479. PubMed PMID: 2011243.

53. Wilson RS, Boyle PA, Yu L, Barnes LL, Sytsma J, Buchman AS, Bennett DA, Schneider JA. Temporal course and pathologic basis of unawareness of memory loss in dementia. Neurology. 2015;85(11):984-91. Epub 2015/08/28. doi: 10.1212/WNL.0000000000001935. PubMed PMID: 26311746; PMCID: PMC4567465.

54. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273-97.

55. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18-22.

56. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis. 2002;38(4):367-78.

57. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM.

58. Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. 2015;5:10576. Epub 2015/05/28. doi: 10.1038/srep10576. PubMed PMID: 26015273; PMCID: PMC4444969.

59. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature genetics. 2016;48(2):214.

60. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a python library for model selection and hyperparameter optimization. Computational Science & Discovery. 2015;8(1):014008.

61. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC bioinformatics. 2010;11(1):587.

62. Wingo TS, Kotlar A, Cutler DJ. MPD: multiplex primer design for next-generation targeted sequencing. BMC bioinformatics. 2017;18(1):1-5.

63. Das S, Abecasis GR, Browning BL. Genotype imputation from large reference panels. Annu Rev Genomics Hum Genet. 2018;19(1):73-96.

64. Abecasis GR, Ghosh D, Nichols TE. Linkage disequilibrium: ancient history drives the new genetics. Human heredity. 2005;59(2):118-24.

65. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome biology. 2011;12(1):1-13.

66. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA. DNA methylation profiling of human chromosomes 6, 20 and 22. Nature genetics. 2006;38(12):1378-85.

67. Yao DW, O'Connor LJ, Price AL, Gusev A. Quantifying genetic effects on disease mediated by assayed gene expression levels. Nature Genetics. 2020:1-8.

68. Rosenthal SL, Barmada MM, Wang X, Demirci FY, Kamboh MI. Connecting the dots: potential of data integration to identify regulatory SNPs in late-onset Alzheimer's disease GWAS findings. PloS one. 2014;9(4):e95152.

69. Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, Younkin SG, Saykin AJ, Sweet RA, Feingold E. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Molecular psychiatry. 2012;17(12):1340-6.

70. Altuna M, Urdánoz-Casado A, de Gordoa JS-R, Zelaya MV, Labarga A, Lepesant JM, Roldán M, Blanco-Luquin I, Perdones Á, Larumbe R. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. Clinical epigenetics. 2019;11(1):91.

71. Johnson NE. Myotonic muscular dystrophies. CONTINUUM: Lifelong Learning in Neurology. 2019;25(6):1682-95.

72. Tavares IA, Touma D, Lynham S, Troakes C, Schober M, Causevic M, Garg R, Noble W, Killick R, Bodi I. Prostate-derived sterile 20-like kinases (PSKs/TAOKs) phosphorylate tau protein and are activated in tangle-bearing neurons in Alzheimer disease. Journal of Biological Chemistry. 2013;288(21):15418-29.

73. Hall M, Quignodon L, Desvergne B. Peroxisome proliferator-activated receptor β/δ in the brain: facts and hypothesis. PPAR research. 2008;2008.

74. Braissant O, Foufelle F, Scotto C, Dauça M, Wahli W. Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha,-beta, and-gamma in the adult rat. Endocrinology. 1996;137(1):354-66.

75. Shin HD, Park BL, Kim LH, Jung HS, Cho YM, Moon MK, Park YJ, Lee HK, Park KS. Genetic polymorphisms in peroxisome proliferator-activated receptor δ associated with obesity. Diabetes. 2004;53(3):847-51.

76. Trejo J, Massamiri T, Deng T, Dewji NN, Bayney RM, Brown JH. A direct role for protein kinase C and the transcription factor Jun/AP-1 in the regulation of the Alzheimer's beta-amyloid precursor protein gene. Journal of Biological Chemistry. 1994;269(34):21682-90.

77. Johnson EC, Dammer EB, Duong DM, Ping L, Zhou M, Yin L, Higginbotham LA, Guajardo A, White B, Troncoso JC. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nature medicine. 2020;26(5):769-80.

78. Schneider JA, Aggarwal NT, Barnes L, Boyle P, Bennett DA. The neuropathology of older persons with and without dementia from community versus clinic cohorts. Journal of Alzheimer's Disease. 2009;18(3):691-701.

79. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic acids research. 2005;33(18):5868-77.

80. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits2012.

81. MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, Abecasis G, Adams D, Altman R, Antonarakis S, Ashley E. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469-76.

82. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics. 2010;11(6):415-25.

83. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research. 2014;42(D1):D1001-D6.

84. Zhu Y, Tazearslan C, Suh Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. Experimental Biology and Medicine. 2017;242(13):1325-34.

85. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Human molecular genetics. 2015;24(R1):R102-R10.

86. Rojano E, Seoane P, Ranea JA, Perkins JR. Regulatory variants: from detection to predicting impact. Briefings in bioinformatics. 2019;20(5):1639-54.

87. Lipscomb CE. Medical subject headings (MeSH). Bulletin of the Medical Library Association. 2000;88(3):265.

88. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA. Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. European Journal of Human Genetics. 2014;22(1):144-7.

89. b GPCCAMGAmwoau, 13 PgBCoMGRADHKCLSLLMDRJWM, MIT BIo, 3 HLESADMGSBGN, 12 EBIFPCLLRSREZ-BX, 8 IBDRGRHSJTKZ, T. UNIoHSS, 2

UoOMGA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

90. Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2008;39(2):539-50.

91. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536-43. Epub 2015/01/15. doi: 10.1093/bioinformatics/btv009. PubMed PMID: 25583119; PMCID: PMC4426838.

92. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome biology. 2016;17(1):1-14.

93. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.

94. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. 2009;22(10):1345-59.

95. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2014;31(5):761-3.

96. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. A method to predict the impact of regulatory variants from DNA sequence. Nature genetics. 2015;47(8):955-61.

97. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-20. Epub 2016/01/05. doi: 10.1038/ng.3477. PubMed PMID: 26727659; PMCID: PMC4731313.

98. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49(4):618-24. Epub 2017/03/14. doi: 10.1038/ng.3810. PubMed PMID: 28288115; PMCID: PMC5395419.

99. Zhou L, Zhao F. Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med. 2018;10(1):53. Epub 2018/07/13. doi: 10.1186/s13073-018-0565-y. PubMed PMID: 29996888; PMCID: PMC6042373.

100. Davis J, Goadrich M, editors. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; 2006.

101. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3):e0118432.

102. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940-1.

103. Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucleic acids research. 2018;46(W1):W109-W13.

104. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nature genetics. 2004;36(5):431-2.

105. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic acids research. 2015;43(D1):D805-D11.

106. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42(D1):D980-D5.

107. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189-96. Epub 2012/11/24. doi: 10.1093/bioinformatics/bts680. PubMed PMID: 23175756; PMCID: PMC3546795.

108. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337(6099):1190-5.

109. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L. The NCBI dbGaP database of genotypes and phenotypes. Nature genetics. 2007;39(10):1181-6.

110. Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics. 2015;47(9):979-86.

111. Vorstman JA, Breetvelt EJ, Thode KI, Chow EW, Bassett AS. Expression of autism spectrum and schizophrenia in patients with a 22q11. 2 deletion. Schizophrenia research. 2013;143(1):55-9.

112. Martínez A, Sanchez-Lopez M, Varadé J, Mas A, Martín MC, de Las Heras V, Arroyo R, Mendoza JL, Díaz-Rubio M, Fernández-Gutiérrez B. Role of the MHC2TA gene in autoimmune diseases. Annals of the rheumatic diseases. 2007;66(3):325-9.

113. Swanberg M, Lidman O, Padyukov L, Eriksson P, Åkesson E, Jagodic M, Lobell A, Khademi M, Börjesson O, Lindgren CM. MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nature genetics. 2005;37(5):486-94.

114. Han J-W, Zheng H-F, Cui Y, Sun L-D, Ye D-Q, Hu Z, Xu J-H, Cai Z-M, Huang W, Zhao G-P. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nature genetics. 2009;41(11):1234-7.

115. Maude SL, Teachey DT, Porter DL, Grupp SA. CD19-targeted chimeric antigen receptor T-cell therapy for acute lymphoblastic leukemia. Blood, The Journal of the American Society of Hematology. 2015;125(26):4017-23.

116. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A. mRNA-Seq whole-transcriptome analysis of a single cell. Nature methods. 2009;6(5):377-82.

117. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486-90.

118. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi–C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268-76.

119. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215-6.

120. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods. 2012;9(5):473-6.

121. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, Diegel M, Dunn D, Neri F, Teodosiadis A. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020;584(7820):244-51.

122. Chen C, Zhang S, Zhang X-S. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. Nucleic acids research. 2013;41(20):9230-42.

123. Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. Nucleic acids research. 2017;45(17):9823-36.

124. Choi H, Fermin D, Nesvizhskii AI, Ghosh D, Qin ZS. Sparsely correlated hidden Markov models with application to genome-wide location studies. Bioinformatics. 2013;29(5):533-41.

125. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics. 2013;14(1):1-14.

126. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology. 2010;28(5):495-501.

127. Xu T, Jin P, Qin ZS. Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. Bioinformatics. 2020;36(3):690-7.

128. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease Ontology: a backbone for disease semantic integration. Nucleic acids research. 2012;40(D1):D940-D6.

129. Zhang C, Chen YE, Zhang S, Li JJ. Information-theoretic classification accuracy: a criterion that guides data-driven combination of ambiguous outcome labels in multi-class classification. arXiv preprint arXiv:210900582. 2021.

130. Vorugunti CS, Gorthi RKS, Pulabaigari V, editors. Online signature verification by few-shot separable convolution based deep learning. 2019 International Conference on Document Analysis and Recognition (ICDAR); 2019: IEEE.

131. Schroff F, Kalenichenko D, Philbin J, editors. Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.

132. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis J, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics. 2008;9(5):356-69.

133. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics. 2011;12(8):529-41.

134. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931-4. Epub 2015/08/25. doi: 10.1038/nmeth.3547. PubMed PMID: 26301843; PMCID: PMC4768299.

135. Yang L, Jin R. Distance metric learning: A comprehensive survey. Michigan State Universiy. 2006;2(2):4.

136. Huang Y, Sun X, Jiang H, Yu S, Robins C, Armstrong MJ, Li R, Mei Z, Shi X, Gerasimov ES. A machine learning approach to brain epigenetic analysis reveals kinases associated with Alzheimer’s disease. Nature communications. 2021;12(1):1-12.

137. Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur). 2020;53(3):1-34.

138. Cao Z, Huang Y, Duan R, Jin P, Qin ZS, Zhang S. Disease category-specific annotation of variants using an ensemble learning framework. Briefings in Bioinformatics. 2022;23(1):bbab438.

139. Oh Song H, Xiang Y, Jegelka S, Savarese S, editors. Deep metric learning via lifted structured feature embedding. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

140. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32.

141. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

142. Method of the Year 2013. Nature Methods. 2014;11(1):1-. doi: 10.1038/nmeth.2801.

143. Teichmann S, Efremova M. Method of the Year 2019: single-cell multimodal omics. Nat Methods. 2020;17(1):2020.

144. Aissa AF, Islam ABMMK, Ariss MM, Go CC, Rader AE, Conrardy RD, Gajda AM, Rubio-Perez C, Valyi-Nagy K, Pasquinelli M, Feldman LE, Green SJ, Lopez-Bigas N, Frolov MV, Benevolenskaya EV. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. Nature Communications. 2021;12(1):1628. doi: 10.1038/s41467-021-21884-z.

145. Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, Gallacher J, Green J, Matthews P, Pell J. UK Biobank: Current status and what it means for epidemiology. Health Policy and Technology. 2012;1(3):123-6.

146. Schork NJ. Genetics of complex disease: approaches, problems, and solutions. American journal of respiratory and critical care medicine. 1997;156(4):S103-S9.

147. Chen J, Tian W. Explaining the disease phenotype of intergenic SNP through predicted long range regulation. Nucleic acids research. 2016;44(18):8641-54.

148. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735-80.

149. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555. 2014.

150. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.