


Disease category-specific annotation of variants using an ensemble learning framework

Zhen Cao[†], Yanting Huang[†], Ran Duan, Peng Jin, Zhaohui S. Qin  and Shihua Zhang 

Corresponding authors: Shihua Zhang, NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. Email: zsh@amss.ac.cn; Zhaohui S. Qin, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA. Email: zhaohui.qin@emory.edu

[†]These authors contributed equally to this work.

Abstract

Understanding the impact of non-coding sequence variants on complex diseases is an essential problem. We present a novel ensemble learning framework—CASAVA, to predict genomic loci in terms of disease category-specific risk. Using disease-associated variants identified by GWAS as training data, and diverse sequencing-based genomics and epigenomics profiles as features, CASAVA provides risk prediction of 24 major categories of diseases throughout the human genome. Our studies showed that CASAVA scores at a genomic locus provide a reasonable prediction of the disease-specific and disease category-specific risk prediction for non-coding variants located within the locus. Taking *MHC2TA* and immune system diseases as an example, we demonstrate the potential of CASAVA in revealing variant-disease associations. A website (<http://zhanglabtools.org/CASAVA>) has been built to facilitate easily access to CASAVA scores.

Key words: complex disease; disease category; functional annotation; non-coding variant; ensemble learning

Introduction

Understanding the role of genetic variants in causing complex diseases is a fundamental problem in genetics [1, 2]. 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 [3]. 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 [4]. 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 [5]. 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 [6, 7]. 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

Zhen Cao was a PhD student at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and now he is a software engineer in the Alibaba Health Information Technology Limited.

Yanting Huang is a PhD student at the Department of Computer Science, Emory University.

Ran Duan is a PhD student at the Department of Software Engineering, Yunnan University.

Peng Jin is a Professor at the Department of Human Genetics, Emory University School of Medicine.

Zhaohui S. Qin is a Professor at the Department of Biostatistics and Bioinformatics, Emory University.

Shihua Zhang is a Professor at the Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

Submitted: 9 August 2021; Received (in revised form): 3 September 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

as essential sources for functional annotation of non-coding variants [8].

Machine learning has been successfully applied to predict the pathogenicity of genetic variants [9–23]. For example, logistic regression was used in CADD that prioritized functional, deleterious, and pathogenic variants [9]. Random forest was used in GWAVA to distinguish disease-implicated variants from benign variants [11]. More information about these methods can be found in a review paper [8].

Despite their popularity, 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 [18].

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 hinges 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 diseases 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.

Materials and methods

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 (Additional file 1: Note S1).

According to the Medical Subject Headings [25] and PheGenI [24], we used 24 representative disease categories (Additional file 2: Table S1). 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 (Additional file 1: Figure S1). 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.

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 [26]. 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 [24]. 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; see Additional file 1: Figure S1). For testing variants, we repeated the sampling procedure ten times.

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 [7, 8]. 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 (Additional file 1: Note S2). 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 2725 genome-wide features, which can be roughly divided into five groups: open chromatin, histone modification, TF binding, gene expression and DNA methylation (Additional file 2: Table S2).

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 2725 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.

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 [27, 28]. The input data are labeled training data (risk variants and benign variants along with their weights). Each variant is represented by 2725 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 [27]. Based on XGBoost, we trained a gradient boosting regression tree classifier using these selected variants [28]. We repeated the under-sampling and training process a number of times (e.g. 100 times) and took their average as our final model (Additional file 1: Note S3). 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 [18] and GWAVA [11]. 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 over-fitting, 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.

Genomic properties of CASAVA scores

We downloaded all the genetic variants from the 1000 Genomes Phase 3 release and predicted these variants using CASAVA scores [26]. According to the Ensemble Variant Effect Predictor [29], 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 [30]. 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 (Additional file 1: Note S4). 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 (Additional file 1: Note S4).

Applying CASAVA to disease-specific risk prediction

We leveraged CASAVA to predict disease-specific risk variants (Additional file 1: Note S5). 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 (Additional file 1: Note S5.5).

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 [31]. 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 (Additional file 1: Note S6).

Comparison with commonly used scoring methods

We compared CASAVA with ten existing functional impact prediction methods: CADD [9], DANN [10], GWAVA [11], FATHMM-MKL [12], GenoCanyon [13], deltaSVM [14], Eigen [16], DIVAN [18], LINSIGHT [19] and PAFA [22]. Though these methods utilized different hypotheses and techniques, they are all reported to be informative of risks of complex diseases (Additional file 1: Note S7). For each method, we downloaded their pre-computed scores and scored the testing variants (Additional file 2: Table S3). 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 [18] and retrained it on the specific diseases we tested.

Performance evaluation

The receiver operating characteristics curve (ROC) is a typical graphical plot that illustrates the classification ability of a binary classifier system [32]. We also considered the precision-recall curve due to the imbalance between risk and benign variants

[33]. 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 [34]. We first evaluated the performance of each method by five-fold cross-validation on training variants (Additional file 1: Note S8). 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.

Case study for immune system diseases

We downloaded all the genetic variants from the 1000 Genomes Phase 3 release [26]. 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 (Additional file 1: Note S9). For all candidate loci, we performed batch query in SNPnexus [35] 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 [36], COSMIC [37], and CinVar [38] databases; to fit the aims of our ensemble classifier, we focused on the query results of the GAD database [36] 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.

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 [28]. We computed the relative importance of each feature using the XGBoost 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 (Additional file 1: Note S10).

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 (Additional file 1: Note S2). 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.

Website for retrieving whole-genome CASAVA scores

To facilitate easy browsing and querying, we designed and implemented a web application for scoring genomic variants, together with an online repository of pre-computed whole-genome CASAVA scores for the 24 disease categories (<http://zhanglabtools.org/CASAVA>). Additionally, we provided an easy-to-use R script for scoring a large number of variants (Additional file 1: Note S11).

Results

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 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 [24] using significance level threshold of P -value 10^{-4} (Figure 1a). We next select corresponding control sets of benign variants from the 1000 Genomes project for each disease category [26]. In the meantime, we collect, curate and process a large set (2725) of genome-wide profiles to be used as features in the classification model (Figure 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 1c).

In the end, CASAVA trains an ensemble classifier for each of the 24 broad disease categories (Figure 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>).

Disease categories

For disease categories, we use those defined by the Medical Subject Headings (MeSH) related to 'diseases' or 'psychiatry and psychology' (Additional file 1: Note S1). The same definition was also used by PheGenI [24, 25]. 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

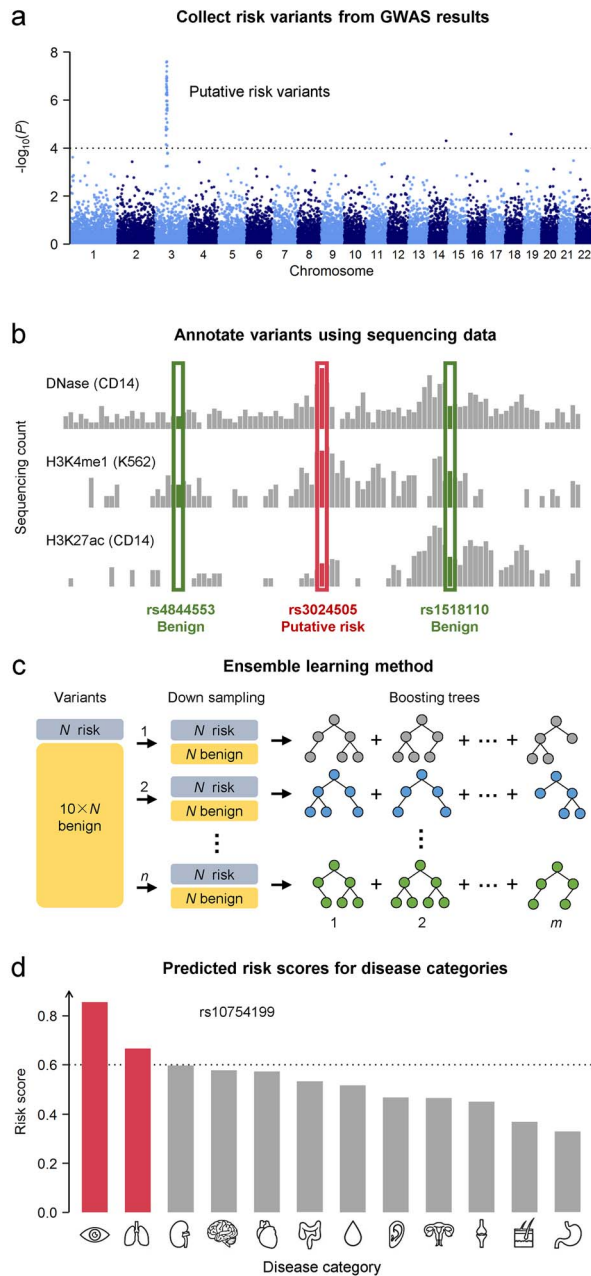


Figure 1. Working pipeline of CASAVA. (a) For each disease category, CASAVA collects known risk variants from existing genome-wide association studies (GWASs) as training data (Additional file 1: Note S1). (b) CASAVA uses genome-wide genomics and epigenomics profiling data as features in its machine learning models (Additional file 1: Note S2). (c) CASAVA applies bagging and boosting techniques to build a classification model for each disease category (Additional file 1: Note S3). (d) CASAVA produces disease category-specific risk prediction for non-coding genetic variants.

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 8065 with a median of 1337 (Additional file 2: Table S1). 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 2995, with a median of 15.

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 [9], DANN [10], GWAVA [11], GenoCanyon [12], FATHMM-MKL [13], deltaSVM [14], Eigen [16], LINSIGHT [19] and PAFA [22]. We found that overall CASAVA performed the best, followed by PAFA and GWAVA in terms of AUC (Additional file 1: Figure S2). 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 (Figure 2a, Additional file 2: Tables S4 and S5). 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 2b); AUPR: 0.35 versus 0.14 (Figure 2c)). Overall, CASAVA performs the best among all methods we compared in terms of AUC and AUPR values (Additional file 1: Figure S2).

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 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 unmatched CASAVA scores (Additional file 1: Note S4.3)? For the first question, we found that specific CASAVA scores (from the corresponding disease) of risk variants are significantly higher (one-tailed Wilcoxon rank-sum test, P -value < 0.05) than those of benign variants (Figure 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 Wilcoxon rank-sum test, P -value < 0.05) than their CASAVA scores from the other 23 disease categories combined (Figure 2d and Additional file 1: Figure S3). These results demonstrated the disease-category specificity in CASAVA score.

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 (Additional file 1: Note S3), bagging and boosting trees [27, 28]. 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 2e, average AUC=0.615).

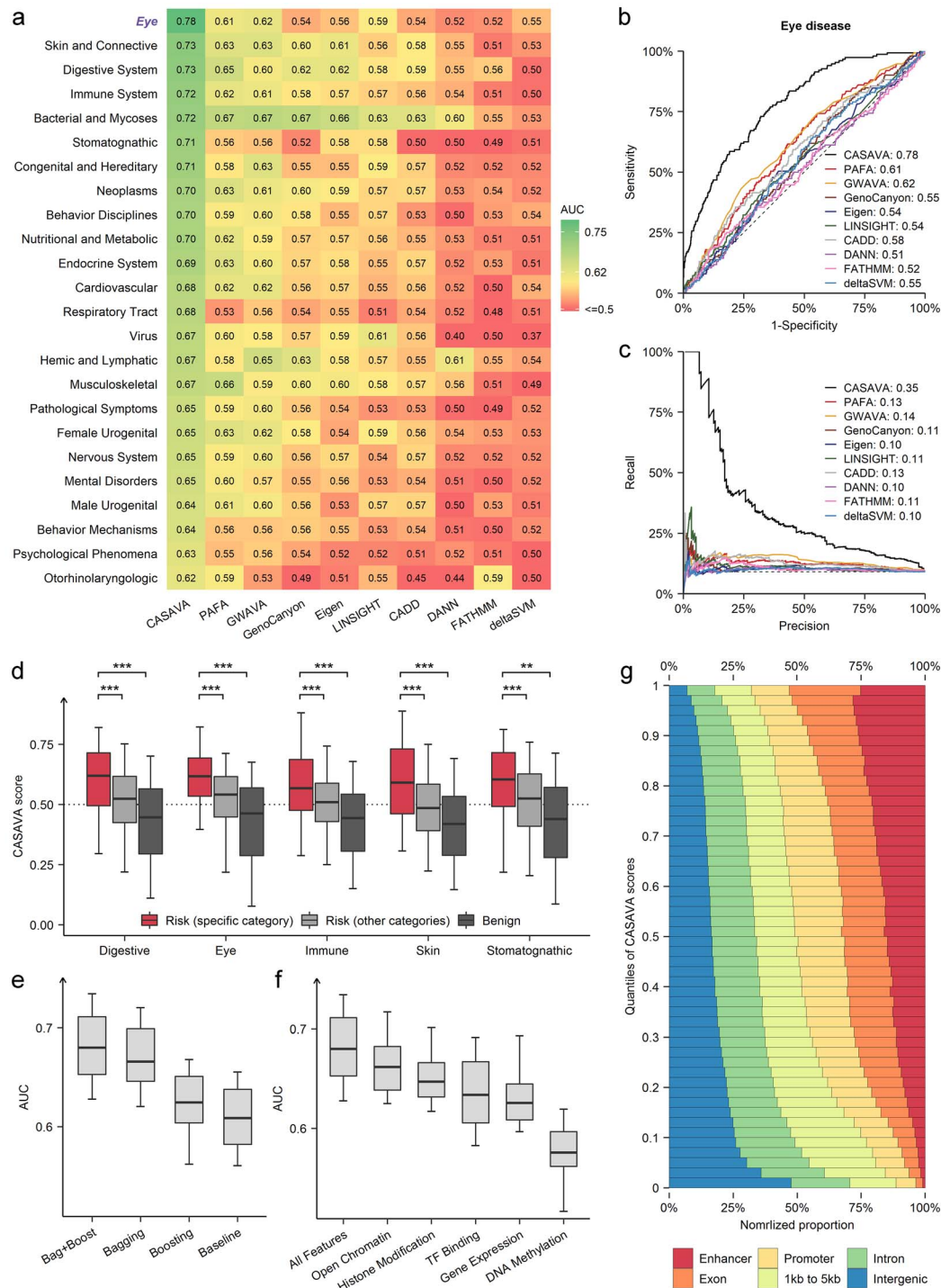


Figure 2. Performance evaluation for disease category-specific risk prediction. (a) Heatmap of AUCs of different methods for 24 disease categories. Each row represents a disease category, and each column represents a scoring method. Some methods have multiple scores, in which case we only use the score with the highest average AUC value (Additional file 1: Note S9). GWAVA score is actually GWAVA TSS-matched score, and Eigen score is actually Eigen-PC score. (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) illustrating 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 (Additional file 1: Note S4). Here we use CASAVA scores for the eye diseases category as an example.

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 (Additional file 1: Figure S4 and Note S3).

Contributions from different groups of features

The current version of CASAVA utilized 2725 features. These features can be divided into five groups: open chromatin, transcription factor (TF) binding, histone modification, DNA methylation and gene expression (Additional file 2: Table S2). 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 2f and Additional file 2: Table S6).

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 (Additional file 2: Table S7). 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.

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 (see Methods) are significantly over-represented (Figure 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 (Additional file 1: Figure S5). 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 [39].

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) [3] and dbGaP [40], 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. Detailed results are summarized in Additional file 1: Figure S6.

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 3a) [41, 42]. In spirit, our strategy is reminiscent of the transfer learning idea that has proved surprisingly effective in many machine learning applications [31].

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 2995 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 3b). 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 (Additional file 2: Tables S8 and S9). 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 (Additional file 1: Note S5.2). CASAVA presented results that are comparable to disease-specific models on the 89 diseases in terms of AUC (Figure 3c, 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

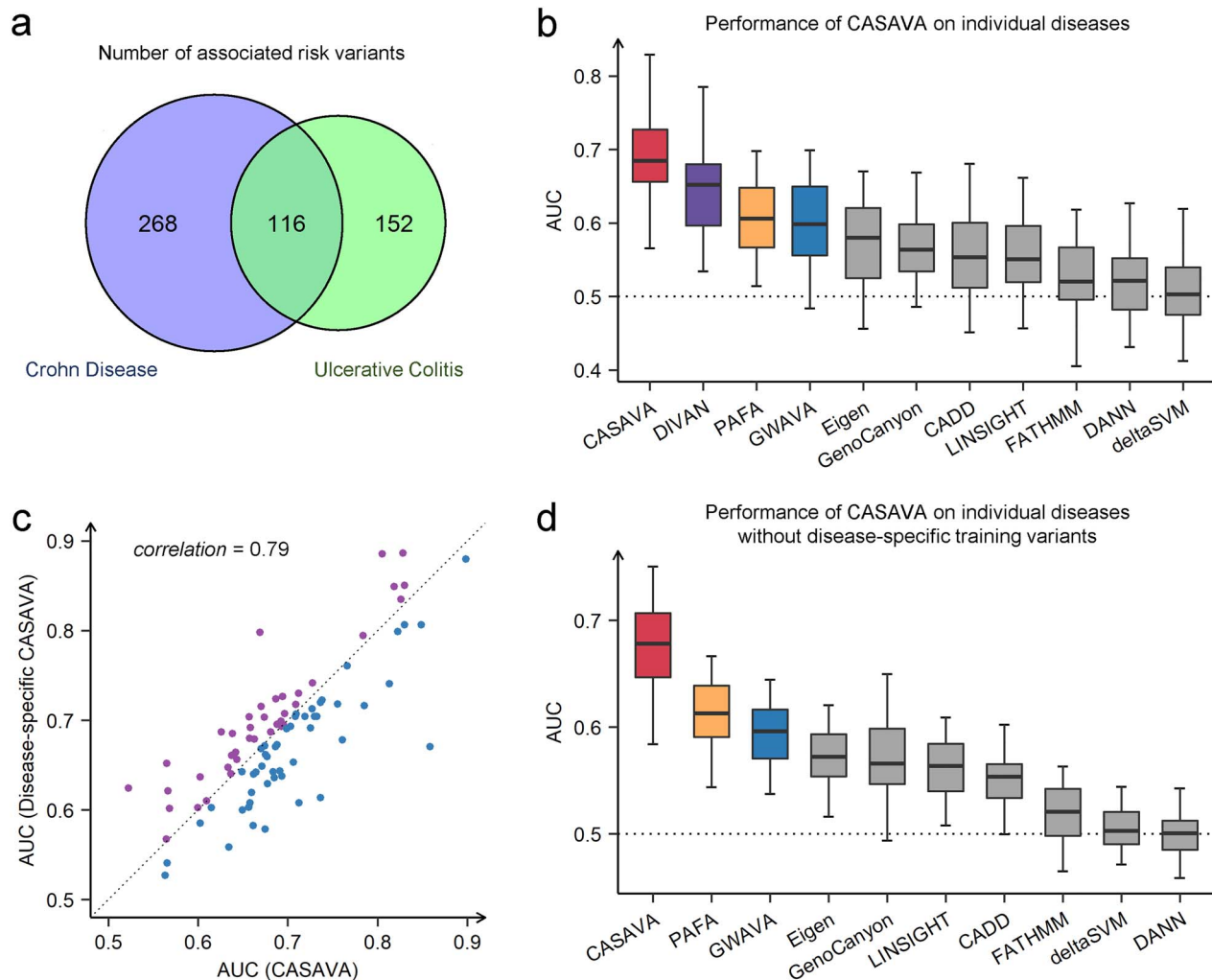


Figure 3. Performance evaluation for disease-specific risk prediction. (a) Venn's diagram for known risk variants that belong to two digestive system diseases. (b) Side-by-side boxplots of 89 AUC values (one for each disease) comparing performance between CASAVA and ten different variant prediction methods. 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 (Additional file 1: Note S6).

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 (Method). 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 (Additional file 1: Note S5.5). Surprisingly, this seemingly simple-minded approach again achieved remarkably better results than existing methods in terms of AUC and AUPR

(Figure 3d, Additional file 2: Tables S10 and S11). 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%).

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 (Additional file 1: Note S6).

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 4a). On average, TrCASAVA lifted the AUC value by 0.013 (Figure 4b, one-tailed paired t-test, $P = 4 \times 10^{-4}$) and the AUPR value by 0.015 (Additional file 1: Figure S7, 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 4c). On average, TrCASAVA lifted the AUC value by 0.005 (Figure 4d, one-tailed paired t-test, $P = 0.04$) and the AUPR value by 0.009 (Additional file 1: Figure S7, 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 (Additional file 1: Note S6), 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 4e). Under the scenario of an extremely small training set, TrCASAVA is likely to significantly improve the prediction results (Figure 4e and Additional file 1: Figure S7). 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 (Figure 4e, 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 (Additional file 1: Figure S7).

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 [43]. 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 [44]. 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 [44, 45]. 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 (Figures 5a and 5b). 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 5c). And for 89 variants out of the 1012 (10.2%), the CASAVA scores corresponding to the immune system diseases not only rank the highest, they are 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, $P = 3 \times 10^{-6}$) [46]. In the gene body region of IKZF1, we found that the average

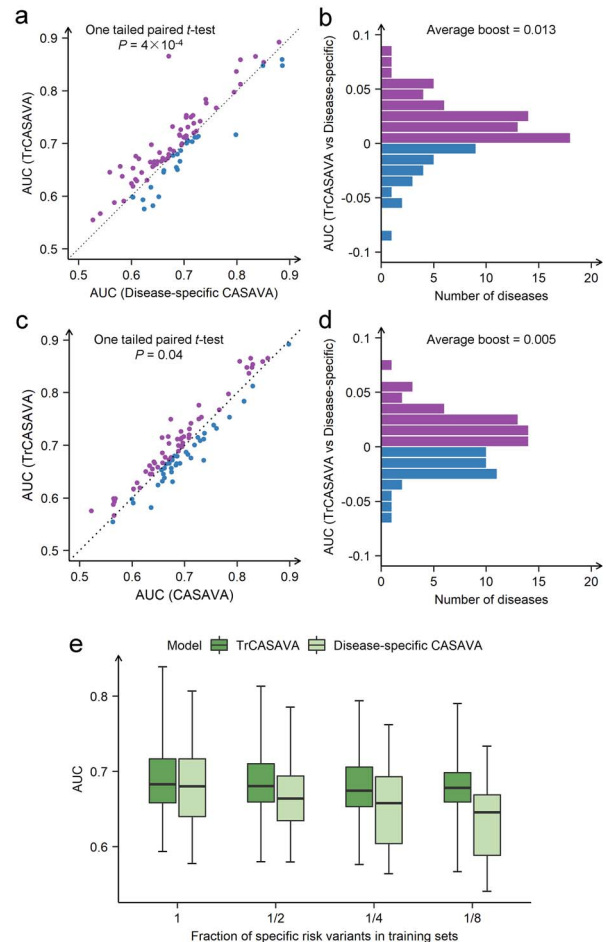


Figure 4. Performance of TrCASAVA for disease-specific risk prediction. (a) Scatter plots comparing AUC values obtained using two different methods: TrCASAVA and the disease-specific version of CASAVA (apply the same ensemble learning framework to each of the 89 diseases). Each point represents one of 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.

score of immune system diseases ranks the highest among all disease categories (Additional file 1: Figure S8). 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, they are 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.

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

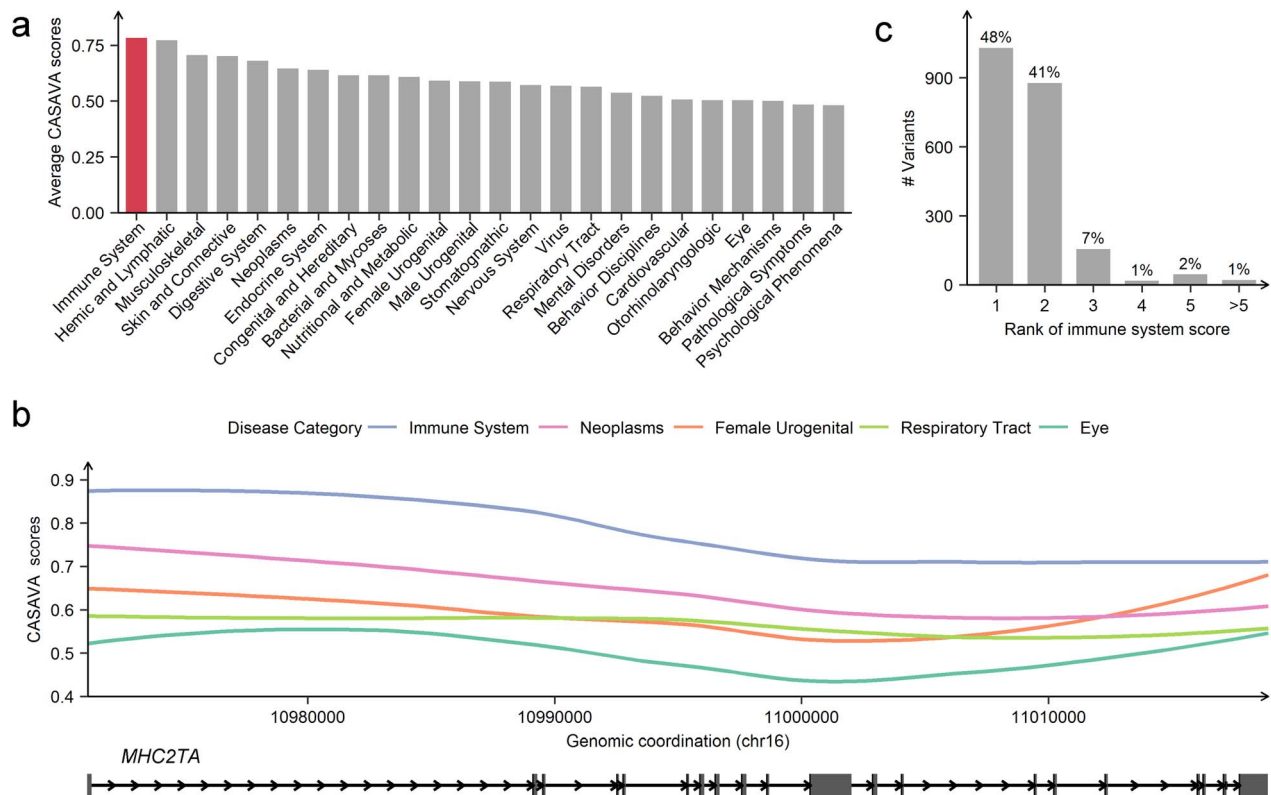


Figure 5. CASAVA identifies *MHC2TA* as an immune disease-related gene. (a) Bar plots of average CASAVA scores inside the gene body region of *MHC2TA*. (b) CASAVA scores inside the gene body region of *MHC2TA*. The CASAVA scores are smoothed using the loess function in R. For representation, only the 1st, 6th, 11th, 16th and 21st highest CASAVA 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.

risk. Overall, we found that features related to histone marks, open chromatin and TF binding contributed more than other types of features (Figure 6a). 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 6b, 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 (Additional file 1: Figure S9). 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 6c). 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 6d), 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 [47]. 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 6e).

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

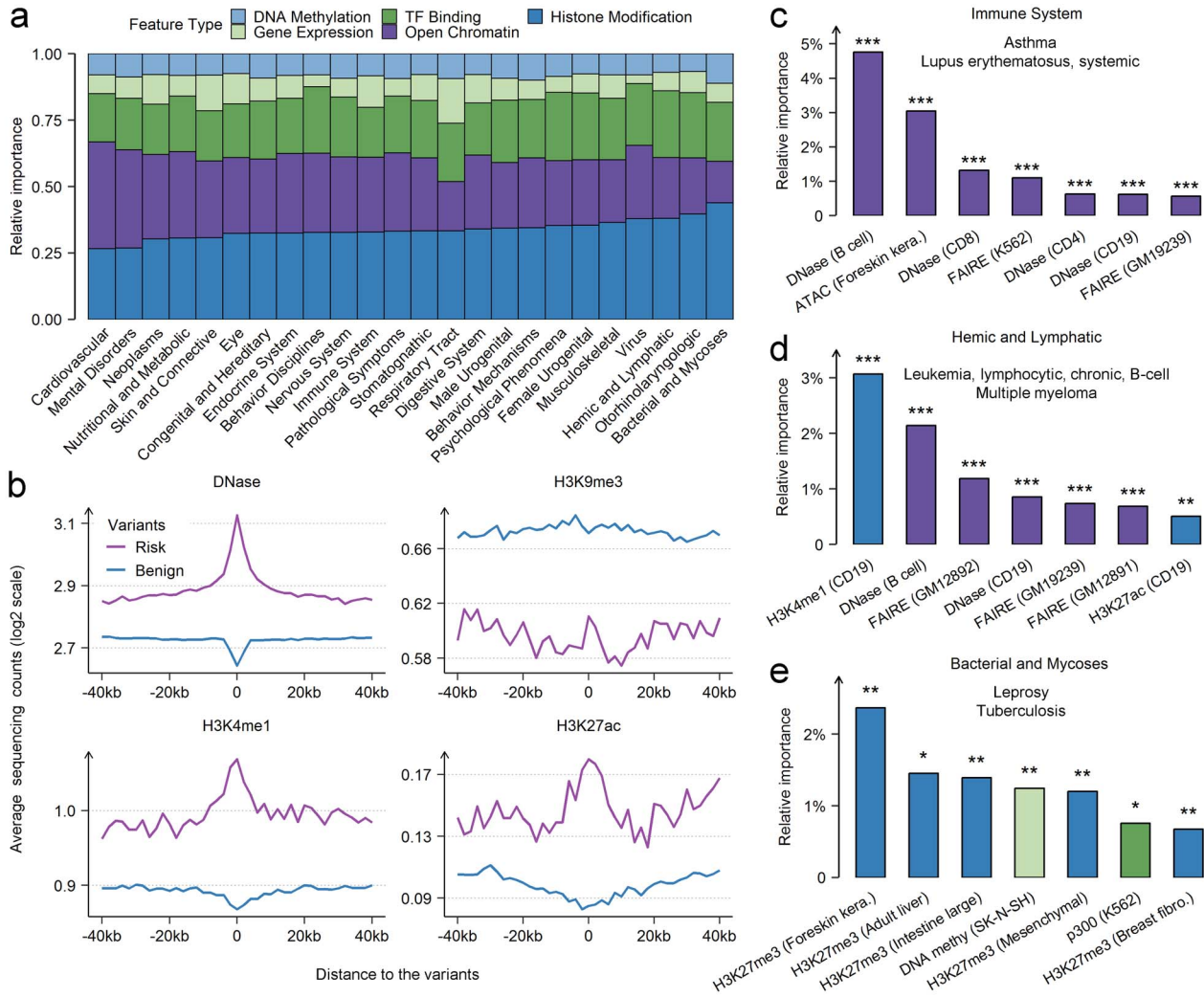


Figure 6. Informative features in CASAVA. (a) Proportions of contribution by different groups of features in the 24 disease category-specific CASAVA models. (b) Selected top-ranked features of the A549 cell line for all risk and benign variants. The line plots show the average read counts (log2 scale) 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. $*1 \times 10^{-8} < P \leq 5 \times 10^{-2}$; $**1 \times 10^{-16} < P \leq 1 \times 10^{-8}$; $***P \leq 1 \times 10^{-16}$.

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 (Additional file 1: Figure S10). 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 2725 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 [48, 49]. We may also use Hi-C data to capture chromatin conformation [50].

We studied the effects of different benign variants for testing (Additional file 1: Note S8.5). 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 (Additional file 1: Figure S11). 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 pose 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 (Additional file 1: Note S8.6): 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 (Additional file 1: Figure S12). However, some methods, such as LINSIGHT [19] or GWAVA TSS-matched score [11], 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 (Additional file 1: Note S4). 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 (Additional file 1: Figure S13). 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 are 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 [51–56], 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 (Additional file 1: Figure S14). 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 [57–59].

In Figure 6a, we notice that variation in the contribution of the five features 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

are following roughly the same order. Additionally, two feature types: DNA methylation and TF binding are relatively stable and the three other features 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) [60]. We will also explore how to combine related diseases or disease categories based on DO for reasonable and sufficient data utilization [61].

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 (Additional file 1: Figure S15). Additionally, we provided pre-computed whole-genome CASAVA scores and an easy-to-use R script for scoring a large number of variants (Additional file 1: Note S11).

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>).

Key Points

- We present a novel ensemble learning framework—CASAVA, to predict genomic loci in terms of disease category-specific risk.
- Compared to nine different competing methods, CASAVA demonstrate the best overall results in terms of both disease category-specific and disease-specific annotation.
- Better results can be achieved when additional known risk variants from related diseases are added under a transfer learning framework. TrCASAVA demonstrate the advantage of pooling together risk variants from similar diseases to boost the performance.
- Using MHC2TA and IKZF1 as examples, we demonstrate that using CASAVA, one could potentially identify novel disease-associated variants or genes.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

This work has been partially supported by the National Key R&D Program of China [2019YFA0709501] to S.Z.; the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [XDPB17] to SZ, the Key-Area Research and Development of Guangdong Province [2020B1111190001] to S.Z., the National Natural Science Foundation of China [61621003] to S.Z.; the National Ten Thousand Talent Program for Young Top-notch Talents to S.Z.; and the CAS Frontier Science Research Key Project for Top Young Scientist [QYZDB-SSW-SYS008] to S.Z.

Authors' contributions

S.Z. and Z.S.Q. designed the study. Z.C., Y.H., P.J., Z.S.Q. and S.Z. developed the method and analyzed the results. R.D. and Z.C. developed the web application. Z.S.Q. and S.Z. supervised the study. Z.C., Y.H., Z.S.Q. and S.Z. wrote the paper.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data Availability

Data used in this research are available from the pre-computed whole-genome CASAVA scores for genetic variants. Official release of CASAVA is available on <https://github.com/zhanglabtools/CASAVA/releases/tag/CASAVA>. All the variants and features can be downloaded from https://zenodo.org/record/4365899#.X_BtZ9gzaUk.

References

1. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014;**508**:469–76.
2. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010;**11**:415–25.
3. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2013;**42**:D1001–6.
4. Zhu Y, Tazearslan C, Suh Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. *Exp Biol Med* 2017;**242**:1325–34.
5. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet* 2015;**24**:R102–10.
6. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 2004;**306**:636–40.
7. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;**28**:1045–8.

8. Rojano E, Seoane P, Ranea JAG, et al. Regulatory variants: from detection to predicting impact. *Brief Bioinform* 2018; **20**:1639–54.
9. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014; **46**:310–5.
10. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015; **31**:761–3.
11. Ritchie GR, Dunham I, Zeggini E, et al. Functional annotation of noncoding sequence variants. *Nat Methods* 2014; **11**: 294–6.
12. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015; **31**:1536–43.
13. Lu Q, Hu Y, Sun J, et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 2015; **5**:10576.
14. Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 2015; **47**:955–61.
15. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015; **12**:931–4.
16. Ionita-Laza I, McCallum K, Xu B, et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016; **48**:214–20.
17. Li MJ, Pan Z, Liu Z, et al. Predicting regulatory variants with composite statistic. *Bioinformatics* 2016; **32**:2729–36.
18. Chen L, Jin P, Qin ZS. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biol* 2016; **17**:252.
19. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 2017; **49**:618–24.
20. Gao L, Uzun Y, Gao P, et al. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat Commun* 2018; **9**:702.
21. Zhou J, Theesfeld CL, Yao K, et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* 2018; **50**:1171–9.
22. Zhou L, Zhao F. Prioritization and functional assessment of noncoding variants associated with complex diseases. *Genome Med* 2018; **10**:53.
23. Chen L, Wang Y, Yao B, et al. TIVAN: tissue-specific cis-eQTL single nucleotide variant annotation and prediction. *Bioinformatics* 2019; **35**:1573–5.
24. Ramos EM, Hoffman D, Junkins HA, et al. Phenotype-genotype integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet* 2014; **22**:144–7.
25. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. *J Am Med Inform Assoc* 2001; **8**:317–23.
26. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**:1061–73.
27. Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst Man Cybern B Cybern* 2009; **39**:539–50.
28. Chen TQ, Guestrin C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016;785–94.
29. Ahn DH, Ozer HG, Hancioglu B, et al. The ensembl variant effect predictor. *Genome Biol* 2016; **17**:122.
30. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature* 2014; **507**:455–61.
31. Pan SJ, Yang Q. A survey on transfer learning. *IEEE T Knowl Data En* 2009; **22**:1345–59.
32. Avis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*. 2006;233–40.
33. 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**:e0118432.
34. Sing T, Sander O, Beerenwinkel N, et al. ROCr: visualizing classifier performance in R. *Bioinformatics* 2005; **21**:3940–1.
35. Dayem Ullah AZ, Oscanoa J, Wang J, et al. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res* 2018; **46**:W109–13.
36. Becker KG, Barnes KC, Bright TJ, et al. The genetic association database. *Nat Genet* 2004; **36**:431–2.
37. Forbes SA, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2015; **43**:D805–11.
38. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 2014; **42**:D980–5.
39. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012; **337**:1190–5.
40. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39**: 1181–6.
41. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015; **47**:979–86.
42. Vorstman JA, Breetvelt EJ, Thode KI, et al. Expression of autism spectrum and schizophrenia in patients with a 22q11.2 deletion. *Schizophr Res* 2013; **143**:55–9.
43. Martínez A, Sánchez-Lopez M, Varadé J, et al. Role of the MHC2TA gene in autoimmune diseases. *Ann Rheum Dis* 2007; **66**:325–9.
44. Swanberg M, Lidman O, Padyukov L, et al. MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. *Nat Genet* 2005; **37**: 486–94.
45. Iikuni N, Ikari K, Momohara S, et al. MHC2TA is associated with rheumatoid arthritis in Japanese patients. *Ann Rheum Dis* 2007; **66**:274–5.
46. Han JW, Zheng HF, Cui Y, et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* 2009; **41**:1234–7.
47. Maude SL, Teachey DT, Porter DL, et al. CD19-targeted chimeric antigen receptor T-cell therapy for acute lymphoblastic leukemia. *Blood* 2015; **125**:4017–23.
48. Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009; **6**:377–82.
49. Buenrostro JD, Wu B, Litzenburger UM, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015; **523**:486–90.

50. Belton JM, McCord RP, Gibcus JH, et al. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 2012;**58**:268–76.
51. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;**9**:215–6.
52. Hoffman MM, Buske OJ, Wang J, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 2012;**9**:473–6.
53. Meuleman W, Muratov A, Rynes E, et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* 2020;**584**:244–51.
54. Chen C, Zhang S, Zhang XS. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. *Nucleic Acids Res* 2013;**41**:9230–42.
55. Zhang Y, Hardison RC. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res* 2017;**45**:9823–36.
56. Choi H, Fermin D, Nesvizhskii AI, et al. Sparsely correlated hidden Markov models with application to genome-wide location studies. *Bioinformatics* 2013;**29**:533–41.
57. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:1–14.
58. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;**28**:495–501.
59. Xu T, Jin P, Qin ZS. Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics* 2020;**36**:690–7.
60. Schriml LM, Arze C, et al. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2012;**40**:940–6.
61. Zhang C, Chen EY, Zhang S, et al. Information-theoretic classification accuracy: a criterion that guides data-driven combination of ambiguous outcome labels in multi-class classification. *Preprint arXiv arXiv* 2021;**2109**:00582.