Published online 12 September 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 21* ***e134***
doi: 10.1093/nar/gkz774

regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants

Shijie Zhang[1,†], Yukun He[1,†], Huanhuan Liu[1], Haoyu Zhai[2], Dandan Huang[1,3], Xianfu Yi[4], Xiaobao Dong[5], Zhao Wang[1], Ke Zhao[1], Yao Zhou[1], Jianhua Wang[1], Hongcheng Yao[6], Hang Xu[6], Zhenglu Yang[7], Pak Chung Sham[8], Kexin Chen[9] and Mulin Jun Li[1,9,*]

[1]Department of Pharmacology, School of Basic Medical Sciences, Tianjin Key Laboratory of Inflammation Biology, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China, [2]Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA, [3]Department of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China, [4]School of Biomedical Engineering, Tianjin Medical University, Tianjin, China, [5]Department of Genetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China, [6]School of Biomedical Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China, [7]College of Computer Science, Nankai University, Tianjin, China, [8]Centre of Genomics Sciences, State Key Laboratory of Brain and Cognitive Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China and [9]Department of Epidemiology and Biostatistics, Tianjin Key Laboratory of Molecular Cancer Epidemiology, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China

Received August 01, 2019; Editorial Decision August 27, 2019; Accepted August 29, 2019

ABSTRACT

Predicting the functional or pathogenic regulatory variants in the human non-coding genome facilitates the interpretation of disease causation. While numerous prediction methods are available, their performance is inconsistent or restricted to specific tasks, which raises the demand of developing comprehensive integration for those methods. Here, we compile whole genome base-wise aggregations, regBase, that incorporate largest prediction scores. Building on different assumptions of causality, we train three composite models to score functional, pathogenic and cancer driver non-coding regulatory variants respectively. We demonstrate the superior and stable performance of our models using independent benchmarks and show great success to fine-map causal regulatory variants on specific locus or at base-wise resolution. We believe that regBase database together with three composite models will be useful in different areas of human genetic studies, such as annotation-based casual variant fine-mapping, pathogenic variant discovery as well as cancer driver mutation identification. regBase is freely available at https://github.com/mulinlab/regBase.

INTRODUCTION

Accurate prediction and prioritization of non-coding regulatory variants are crucial issues in the human genetic studies. Genome-wide association studies (GWASs) have produced numerous single-nucleotide variants (SNVs) that are associated with hundreds of medical traits and diseases, and the majority of the associations are suggested to be mediated by non-coding regulatory codes (1–3). Whole genome sequencing technologies are frequently incorporated into the relevance investigation of non-coding variants in Mendelian disease (4,5), and existing evidence also suggests that non-coding regulatory variants can modulate disease risk by affecting pathogenic coding variant penetrance (6). Given the high volume of disease-causal candidate variants in the regulatory region as well as the expensive downstream functional validations, computationally predicting non-coding regulatory variants has become important and long-standing scientific issue.

In the last few years, a large number of computational methods had been proposed to annotate and predict functional non-coding variants. Building on different predictive assumptions, abundant annotation datasets as well as complementary statistical models, these algorithms have

*To whom correspondence should be addressed. Tel: +86 22 83336668; Fax: +86 22 83336668; Email: mulin0424.li@gmail.com
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

achieved great successes to prioritize functional, pathogenic and cancer-relevant non-coding regulatory variants (7–10). However, the state-of-the-art benchmarks showed poor concordance among the prediction scores of several existing methods (11–13). To comprehensively evaluate the regulatory potential or pathogenesis of certain SNV outside the protein-coding region, researchers now have to collect and compare scores from different resources, even need to download huge pre-computed files or manually calculate prediction scores. The overwhelming growth of new prediction tools further complicates such retrieval processes. In addition, the incomplete understanding and the functional complexity of regulatory DNA impede the development of single but versatile model that is able to accurately predict causal regulatory variants affecting different biological processes. For example, recent commonly adopted algorithms that integrate evolutionary constraint, epigenomics and sequence features, such as CADD (14,15), GWAVA (16), FunSeq2 (17) and fitCons (18), usually achieved limited predictive power for expression-modulating variants from *in vivo* saturation mutagenesis of an enhancer (19), or allele imbalanced variants influence critical molecular traits in the transcriptional regulation, like chromatin accessibility (20). Furthermore, compared with the functional regulatory variants prioritization, it is more challenging to predict pathogenic regulatory variants that underlie the development of Mendelian disorders or cancers (5,21). The insufficient accumulation of known pathogenic regulatory variants largely inhibits the characterization of their key discriminative features that is different from disease-free regulatory mutations.

In this work, we comprehensively integrate non-coding variant prediction scores from 23 tools for base-wise annotation of human genome, called regBase. As such, regBase provides first-time convenience to prioritize functional regulatory SNVs and to assist the fine mapping of causal regulatory SNVs without queries from numerus resources. Inspired by the evident significance of ensemble prediction for pathogenic/deleterious nonsynonymous substitution, we systematically construct three composite models to score functional, pathogenic and cancer driver non-coding regulatory SNVs. We illustrate the discriminatory abilities and applicable scenarios of the proposed models by independent datasets and case studies. regBase and associated models are freely available for download at https://github.com/mulinlab/regBase.

MATERIALS AND METHODS

Collecting, processing and integrating functional scores for non-coding regulatory variants

We downloaded base-wise precomputed scores for almost all possible substitutions of single nucleotide variant (SNV) in the human reference genome from 13 existing tools, including CADD (14,15), CDTS (22), CScape (23), DANN (24), Eigen (25), FATHMM-MKL (26), FATHMM-XF (27), FIRE (28), fitCons (18), FunSeq2 (17), GenoCanyon (29), LINSIGHT (30) and ReMM (31). We called this aggregated resource as regBase. For tool score recorded by interval-level value, such as CDTS, fitCons and LINSIGHT, we transformed continuous position into base-wise position and assigned the same score. Since some tools only support functional annotations for 1000 Genomes Project variants (32) or are inefficient to compute variant scores, we collected or generated functional scores of additional 10 tools for only biallelic variants from 1000 Genomes Project phase 3, including Basset (33), CATO (20), DanQ (34), DeepSEA (35), deltaSVM (36), FunSeq (37), GWAS3D (37), GWAVA_TSS (16), RSVP (38) and SuRFR (39) (see Supplementary Tables S1 and S2 for details). We extracted 1000 Genomes Project biallelic variants from 13 base-wise precomputed scores and merged together with above 10 scores to generate a database that contains 23 tools for all biallelic variants, called regBase Common. Missing score values were replaced with '.' and genomic position of all variants were based on GRCh37/hg19. We also ranked all scores in each set and normalized them by PHRED-scaled score ($-10*\log_{10}(\text{rank}/\text{total})$). The integrated database is tab delimited and indexed by Tabix (40).

Correlation analysis

Three benchmark datasets were incorporated to evaluate the prediction consistency of existing tools including (i) the Human Gene Mutation Database (HGMD) functional regulatory variants used by GWAVA (41); (ii) the ClinVar (201812 release) regulatory variants (42) with 'CLNSIG = Pathogenic or CLNSIG = Benign' and only obtaining non-coding attributes by VEP (43) (not including splicing-altered consequences); (iii) expression-modulating variants identified by massively parallel reporter assay (MPRA) with more than 1.5 $\log_2$ fold expression level change between alleles (44). Pearson correlation test and hierarchical clustering were used to evaluate the relationships of integrated tools upon these non-coding regulatory variant datasets, in which variants with missing value for any tools will be excluded (Supplementary Table S3).

Construction of training dataset

We designed three training datasets to predict different categories of functional non-coding regulatory variants as follows:

regBase_REG and regBase_REG_Common dataset: assuming to functional regulatory variants regardless of functional direction and pathogenicity. We used our previously compiled functional regulatory variants dataset in PRVCS (11), which integrates four different resources including (i) the HGMD public dataset used by GWAVA; (ii) the ClinVar pathogenic variants in the non-coding region compiled by GWAVA; (iii) validated regulatory variants from the OregAnno database (45); (iv) fine-mapped disease-causal regulatory SNPs for 39 immune and non-immune diseases (46). Since some existing tools can only calculate prediction scores for known germline variants in the human population, to incorporate as many scores as possible and avoid missing values for very rare/*de novo*/somatic variants, we only kept variants which appear in the 1000 Genomes project. Negative controls were sampled from allele frequency matched non-coding variants in the independent linkage disequilibrium (LD) with positive variants from 1000 Genomes Project.

regBase_PAT dataset: assuming to pathogenic regulatory variants. We incorporated ClinVar (201812 release) pathogenic regulatory mutations with 'CLNSIG = Pathogenic' and only kept the mutations in the non-coding region by VEP annotations (not including splicing-altered consequences). We also included regulatory Mendelian mutations in the non-coding region from Genomiser (31) and merged with ClinVar data. For negative dataset, we randomly drew benign mutations with 'CLNSIG = Benign' from ClinVar, and used the same strategy to retain non-coding mutations.

regBase_CAN dataset: assuming to cancer recurrent regulatory somatic mutations. For positive set, we downloaded COSMIC v84 non-coding mutations and selected ones having recurrence rate ≥ 10. For negative set, we sampled private non-coding somatic mutations with recurrence = 1 and PhyloP = 0 (47) (see Supplementary Table S4 for variant statistics).

Gradient Tree Boosting model and evaluation

We made use of Gradient Tree Boosting (GTB) algorithm in our predictive model. In general, GTB is a special form of Gradient Boosting Machine, which makes prediction by combining the results of multiple weak learners, typically decision tree. We used XGBoost classifier as the implementation of GTB algorithm. XGBoost is a scalable end-to-end tree boosting system and has achieved the state-of-art performance in plenty of tasks (48). Its sparsity-aware split finding makes it suitable for the task as missing value was commonly appeared in our datasets. We performed grid search based on 10-fold cross-validation on training set in order to tune the hyper-parameters. While tuning training datasets with the unbalanced positive and negative samples, we adjusted the weight of positive samples according to the ratio of two classes. Receiver operating characteristic (ROC) curve and area under the receiver operating characteristics curve (AUC) were used to evaluate the performance of model during grid search. We also compared XGBoost algorithm with other machine learning algorithms including SVM, AdaBoost and RandomForest. Feature contribution was measured by permutation importance and SHapley Additive exPlanation (SHAP) approaches (49). Pearson correlation test and hierarchical clustering were used to evaluate the correlation between our proposed scores under four models with different training datasets and existing prediction scores.

Construction of independent testing datasets

We assembled eight independent testing datasets that were not used to train almost all of existing tools and our combined models, including (i) Brown_eQTL dataset: 11 tissue/cell type-specific eQTLs fine-mapping data that was profiled by Brown and colleagues (50). To further acquire more significant eQTL SNPs, we applied $\log_{10}$BF cutoff values of 10% FDR for each tissue/cell type; (ii) GTEx_eQTL dataset: GTEx V6 44 tissues-specific eQTLs within CAVIAR (51) 95% fine-mapped credible set from UCSC (52); (iii) GWAS_5E-8 dataset: GWAS disease-associated regulatory variants with *P*-value < 5E–8 from GWAS Catalog v1.0.1 (53); (iv) GWAS_1E-5 dataset: GWAS disease-associated regulatory variants with *P*-value < 1E–5 from GWAS Catalog v1.0.1 (53); (v) Somatic_eQTL dataset: recurrent somatic mutations from COSMIC V84 with recurrence ≥ 2 within significant flanking intervals per somatic eGene (54); (vi) Rare_Patho_SNV dataset: high confidence pathogenic regulatory variants curated by two recent publications. These variants were recorded to cause Mendelian diseases with different levels of evidence (22,55); (vii) ASD_denovo_SNV dataset: experimentally validated transcriptional-regulation-disruption *de novo* mutations associated with autism spectrum disorder (ASD) (56); 8) MPRA_eQTL dataset: significant expression modulating variants ($\log_2$FC > 1.5) by MPRA in lymphoblastoid cell lines (44). We also generated corresponding controls for above datasets using different sampling strategies. For Brown_eQTL and GTEx_eQTL dataset, we randomly sampled allele frequency matched non-coding variants in the 10 kb transcription start site (TSS) regions of randomly selected genes. For GWAS_5E–8 and GWAS_1E–5 dataset, we sampled allele frequency matched non-coding variants in the independent LD with positive variants from 1000 Genomes Project. For Somatic_eQTL dataset, we sampled private non-coding somatic mutations from COSMIC V84 with recurrence = 1 and PhyloP = 0. For Rare_Patho_SNV dataset we used non-coding benign variants from ClinVar (CLNSIG = Benign, 201812 release–201907 release). For ASD_denovo_SNV dataset, we sampled nearest non-coding non-pathogenic *de novo* mutations in the siblings of ASD patients. For MPRA_eQTL dataset, we used nonexpression-modulating variants ($\log_2$FC < 0.005) by MPRA in lymphoblastoid cell lines. Importantly, we excluded all positive and negative samples that have been incorporated in our training datasets. For Rare_Patho_SNV and ASD_denovo_SNV, we also removed samples which had been recorded in the HGMD database (see Supplementary Table S5 for statistics of these testing datasets).

MPRA model and evaluation

Additional regBase_MPRA and regBase_MPRA_Common model were trained on MPRA_eQTL dataset and evaluated by 10-fold cross-validation. We also collected MPRA positive variants from three publications (56–58) and constructed an independent MPRA_intergrated_SNV testing dataset. Negative dataset was sampled from allele frequency matched non-coding variants in the 10 kb TSS regions of randomly selected genes.

Benchmark schemes

We compared our composite models with integrated tools and two existing ensemble methods (PRVCS (11) and IW-Scoring (12)) using above six independent testing datasets. Positive predictive values (PPV), negative predictive values (NPV), false positive rate (FPR), false negative rate (FNR), sensitivity, specificity, accuracy, precision, recall, F1 score and Matthews correlation coefficient (MCC) were calculated according to Maximal Youden's index during the measurement of ROC and AUC. We also calculated the correlation between true labels and prediction scores for each evaluation using Pearson correlation test.

Causal variants prioritization for 5p15.33 TERT region

We collected significant trait/disease associated SNPs from GWAS catalog (*P*-value < 5E–8) and GWAS fine-mapping results from literatures at the 5p15.33 TERT region (Human GRCh37, chr5:1.22–1.37mb). We used LocusZoom (59) to visualize these disease-associated and fine-mapped SNPs on 1000 Genomes EUR population. To investigate the performance of regBase composite methods for causal variant prioritization, we extracted and normalized the raw scores of all tools in the 5p15.33 TERT region to generate regional PHRED-scaled scores. We further evaluated the sum or distribution of PHRED scores for all collected fine-mapped SNPs across different tools. Since some tools contain equal scores at this region and this will reduce the discrimination of true causal variants, we removed tools that obtain >25% equal scores in the evaluation.

Base-wise evaluation for saturation mutagenesis of *ALDOB* enhancer

We used *in vivo* saturation mutagenesis data for *ALDOB* enhancer to perform base-wise evaluation among our proposed models and existing methods (60). Tools with high missing rate and low uniqueness for 259 bp *ALDOB* enhancer were identified and excluded in following comparison. Pearson correlation coefficient was used to investigate the concordance between prediction scores and true fold changes of experiment.

Discrimination of variant-level pathogenic alleles

We downloaded non-coding pathogenic alleles and matched non-pathogenic human derived alleles from three simulated datasets (13). Briefly, non-coding SNVs with pathogenic alleles never observed in diverse non-human placental mammals were selected, and matched non-pathogenic human derived alleles at the same position were drawn with varied frequencies, which yielded 55 453 (57 mammals and 5–15% derived allele frequency), 47 799 (5–95% derived allele frequency) and 79 506 positions (11 primates) respectively. To ensure a valid evaluation, we discarded prediction tools that frequently predict the same score between simulated pathogenic and non-pathogenic alleles. We calculated *Z*-score for each allele and prioritized the distance of paired Z-score for each variant position.

RESULTS

Generally, this work consists of four major parts, including (i) integration of whole genome base-wise prediction scores; (ii) construction of composite prediction models; (iii) model evaluation using independent testing datasets; (iv) application of established models for causal regulatory variants identification. The study workflow was shown in Figure 1A.

Base-wise aggregation of non-coding regulatory variant prediction scores

We processed and compiled an integrative resource for prediction scores from 23 different tools on functional annotation of non-coding variants, including Basset (33), CADD (14,15), CATO (20), CDTS (22), CScape (23), DANN (24), DanQ (34), DeepSEA (35), deltaSVM (36), Eigen (25), FATHMM-MKL (26), FATHMM-XF (27), FIRE (28), fitCons (18), FunSeq (37), FunSeq2 (17), GenoCanyon (29), GWAS3D (37), GWAVA (16), LINSIGHT (30), ReMM (31), RSVP (38) and SuRFR (39) (Supplementary Table S1). Since some tools only support annotations for 1000 Genomes Project variants (32), or take long runtime to compute functional scores, we first built a database, called regBase Common, which contains functional scores from 23 tools for 38 248 779 in the 1000 Genomes Project phase 3. Among these integrated datasets, 13 tools provide precomputed scores for almost all possible substitutions of SNV in the human reference genome. Therefore, we also constructed a complete base-wise aggregation of non-coding variant functional scores for 8 575 894 770 substitutions of SNV (same with CADD pre-calculated alleles which consists of all possible substitutions in the human reference genome GRCh37), called regBase (Supplementary Table S2). We summarized the missing values in our integrated resources, and found that most of tools had less than 2% missing values across the whole genome. However, CATO (65.88%), SuRFR (33.91%) and CDTS (9.64%) exhibited relatively high or moderate missing rates in the regBase Common, and CDTS (13.02%) showed moderate missing rate in the regBase (Supplementary Tables S6 and S7). To facilitate the efficient retrieve and comparison of functional scores of different alleles across tools, we indexed the whole dataset and used a PHRED-scaled method to normalize the raw score of each tool. The regBase and regBase Common can be downloaded from https://github.com/mulinlab/regBase.

Correlation analysis of existing algorithms

Existing non-coding variants prediction algorithms dealt with different predictive objectives and assumptions, which could lead to inconsistent prediction on various application scenarios. To comprehensively evaluate the predictive concordance among our collected scores, we prepared three benchmark datasets that incorporate different pathogenicity/regulatory causality assumptions of non-coding regulatory variants (Supplementary Table S3): (i) functional regulatory variants from the public Human Gene Mutation Database (HGMD) (41) used by GWAVA; (ii) pathogenic and benign regulatory variants from the ClinVar database (42); (iii) experimentally validated expression quantitative trait loci (eQTL) variants from a massively parallel reporter assay (MPRA) (44). Pearson correlation analysis of regBase Common integrated functional scores showed both shared and distinct patterns on these benchmark datasets (Figure 1B). Algorithms trained on similar positive/negative data and features had relatively high pairwise correlations, like DeepSEA and DanQ (Pearson correlation coefficients, $R > 0.7$), or CADD and DANN ($R > 0.6$), or FunSeq and FunSeq2 ($R > 0.5$). However, the majority of tools exhibited weak pairwise correlations ($R < 0.4$) in these regulatory variant datasets, which could be explained by the different training data and features, as well as the various learning models used. Among these tested non-coding regulatory variant datasets, we found the

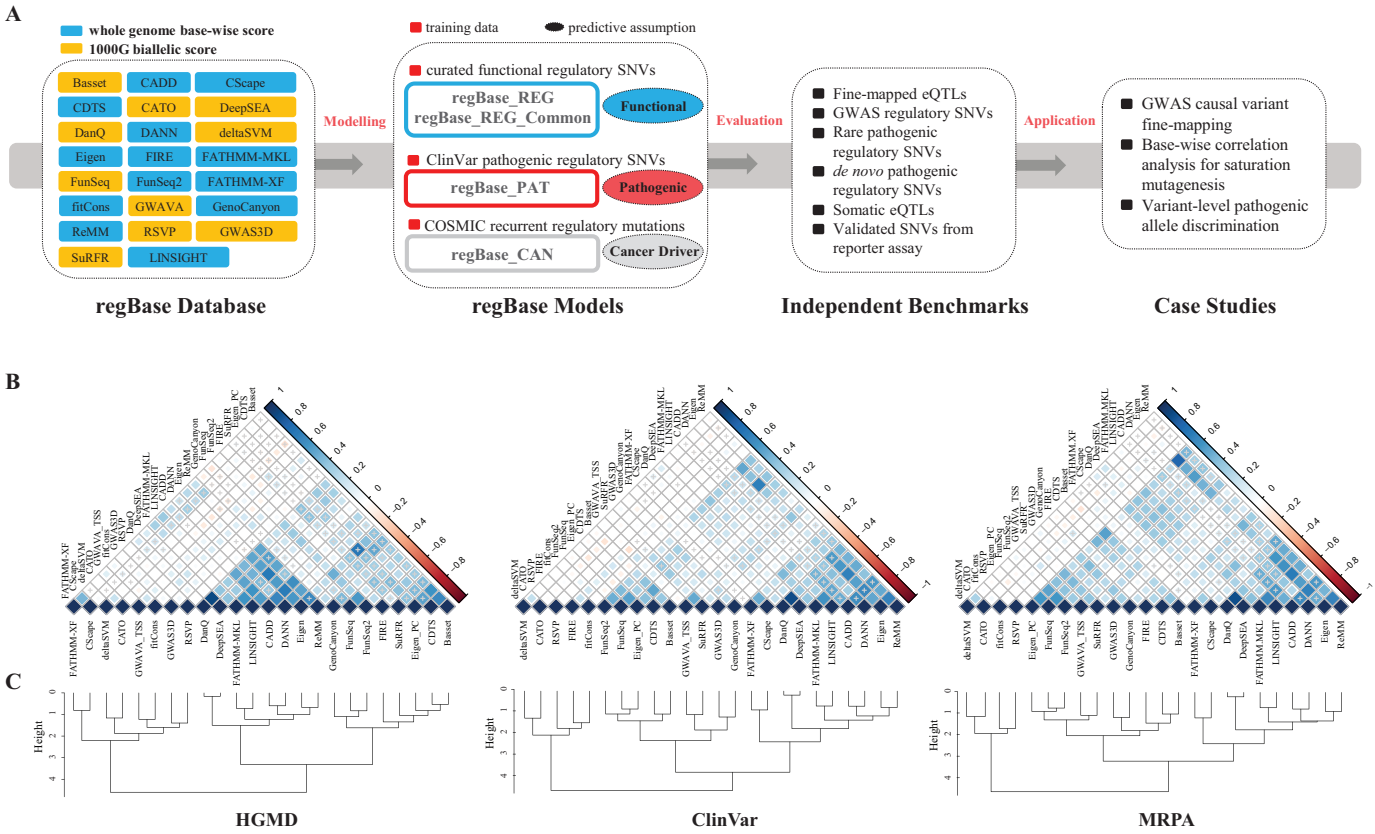

Figure 1. Study workflow and correlation analysis of prediction score among 23 regBase Common integrated tools. (**A**) A flowchart showing the workflow of our regBase study. (**B**) Pearson correlation of 23 regBase Common integrated functional scores on three known functional/pathogenic regulatory variant datasets. Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the square are proportional to the correlation coefficients. Non-significant *P*-value (>0.05) is marked with a cross. (**C**) Hierarchical clustering of regBase Common integrated tools on three known functional/pathogenic regulatory variant datasets. HGMD, the Human Gene Mutation Database functional regulatory variants dataset; ClinVar, the ClinVar pathogenic and benign regulatory variants dataset; MPRA, the expression-modulating variants dataset identified by massively parallel reporter assay.

overall pairwise correlation for MPRA dataset was generally higher than those from other two datasets, implying that current tools may obtain better concordance in eQTL-associated regulatory variant prediction. Since some tested variants were not incorporated or obtained missing values in the regBase Common database, we also performed correlation analysis on 13 complete scores in the regBase database and found similar correlation patterns (Supplementary Figure S1).

To visualize underlying relationships among these tools, we clustered the functional scores according to three above regulatory/pathogenic variant datasets. We found these tools could be generally partitioned into two major subsets, in which each member at the first subset barely associated with other tools within or outside this subset, while members at the second subset were usually correlated with each other (Figure 1C). This result indicates that some tools may capture the unique and important features that is able to distinguish regulatory variants from neutral ones. For example, deltaSVM and CATO learn classification models based on SNV disrupting DNase I hypersensitive site (DHS), and RSVP identifies many informative predictors from gene expression annotations. Interestingly, besides the tools that use exactly same training data or features, we found several tool pairs consistently clustered together in all three results, such as deltaSVM and CATO both utilize variants at DHS as training data. FATHMM-XF co-occured with CScape in the clustering, probably due to their use of similar negative samples and functional annotation features. (Figure 1C and Supplementary Figure S1). To summarize together, our results indicate that the existing non-coding variant functional scoring tools will produce inconsistent predictions across pathogenic/regulatory and neutral variants, and may capture various attributes of functional regulatory codes, suggesting the necessity and importance of systematic integration.

Composite predictions of functional, pathogenic and cancer driver non-coding regulatory variant

Few ensemble prediction models for non-coding regulatory variants were proposed previously. These models only integrated limited number of tools and achieved mediocre performance on pathogenic regulatory variant prediction, especially for predicting somatic regulatory mutation associated with the development of cancer. Given the functional complexity and insufficient accumulation of causal regulatory variants, it is difficult to establish a well-rounded

model that can predict all types of regulatory variants in the current stage. We hence partitioned the non-coding regulatory variant prediction task into three categories, including (i) predicting variant regulatory potential regardless of its functional direction and pathogenicity; (ii) predicting disease-causal regulatory variant; (iii) predicting cancer driver regulatory mutation. Correspondingly, we constructed three independent training datasets (Supplementary Table S4), including (a) functional regulatory variants dataset from our previous PRVCS (11) (regBase_REG); (b) pathogenic regulatory variants dataset from ClinVar and Genomiser (regBase_PAT); (c) highly recurrent regulatory somatic mutations dataset from COSMIC (regBase_CAN). For each positive set, we sampled constrained control set based on the best of our knowledge to alleviate biases (see Materials and Methods for details).

Owing to the potential complementarity and uniqueness of existing non-coding regulatory variant prediction algorithms, we hypothesized that combining functional scores from multiple tools would boost the prediction performance for each aforementioned regulatory variant category. Using the compiled golden standards and regBase scores, we trained three composite models by Gradient Tree Boosting (GTB). We adapted XGBoost classifier as the implementation of GTB algorithm (48), because sparsity-aware split finding of XGBoost make it suitable for the task as missing value are commonly appeared in our regBase features. As all training variants of regBase_REG came from 1000 Genomes Project, we were able to train additional model using regBase Common features (regBase_REG_Common). We tuned the model hyper-parameters by 10-fold cross-validation and evaluated the model performance by receiver operating characteristic (ROC) curve and area under the curve (AUC).

The new composite models significantly improved the prediction performance of the best single tool by 5–22% (Figure 2). Specifically, for functional non-coding regulatory variant prediction, regBase_REG_Common model received average AUC of 0.93 (Figure 2A) and regBase_REG model got 0.89 (Figure 2B). GenoCanyon is always the best single tool with AUC of 0.84 in these two models compared to an average score less than 0.75 achieved by the majority of tools, which implies that integrating more tools with weak but complementary ability could increase the performance of ensemble prediction model. For pathogenic non-coding regulatory variant prediction, regBase_PAT model reached an average AUC of 0.90 (Figure 2C) that exceeds the best tool ReMM by 6% (AUC of 0.84). Remarkably, Tools without training on any ClinVar data, like Eigen, LINSIGHT and CADD, can achieve a comparable performance (AUC $>$ 0.8) with ReMM on predicting disease-causal regulatory variants. This may highlight that evolutionary information and unbiased leaning strategy frequently used in these tools, could be very useful to discriminate mutation pathogenicity or deleteriousness from neutral signals. For the prediction of cancer driver non-coding regulatory mutation, our regBase_CAN model got an unexpectedly high average AUC of 0.91 (Figure 2D) that outperformed the best tool FIRE by 22% (AUC of 0.69). We found most existing algorithms were not specially designed to prioritize somatic regulatory variants except for FunSeq2 and CScape in the regBase database. The preliminary understanding of regulatory codes in the cancer genome and the limited number of cancer driver non-coding variants could be keypoints that inhibited the development of effective prediction model. However, by compositing the effect of existing regulatory variant scoring scheme, we provided an alternative strategy to prioritize non-coding regulatory mutation with cancer driver potential. It is worth noting that some tools received very low or unnormal AUC in above benchmarks, which could be attributed to the discordant predictive assumption with corresponding training dataset.

To investigate the underlying contributions for improved model performance, we first compared the cross-validation results among different machine learning algorithms. We found ensemble learning methods including AdaBoost, RandomForest and XGBoost exhibited better performance than conventional SVM classifier in all training datasets, in which the models trained by XGBoost algorithm showed the best prediction performance (about 2–3% improvements of AUC, Supplementary Figures S2-S5). Second, we estimated the feature importance of our trained XGBoost models and found varied contributions of predictors among them, for instance, GenoCanyon obtained the largest importance in regBase_REG model while CDTS was the best contributor in regBase_PAT model (Supplementary Figure S6). This may imply that the models tend to place higher weight on tools holding similar predictive assumption with corresponding training dataset. Besides the measurement of feature importance, we also used a more interpretable schema, SHAP value, to assess the feature impact on model output. By plotting the SHAP values of every feature for every sample and sorting features by the sum of SHAP value magnitudes over all samples, we found that several features displayed unique SHAP value distribution and may independently contribute to corresponding models, such as GenoCanyon in regBase_REG model and fitCons in regBase_PAT model (Supplementary Figure S7). This further indicates the potential complementarity of collected prediction tools and the necessity of score aggregation. Finally, we performed correlation analysis between our proposed scores under four trained models and existing prediction scores. In general, regBase_REG and regBase_REG_Common models are more correlated with tools used to predict functional regulatory variants, regBase_PAT model is highly correlated with pathogenic variant prediction scores, while regBase_CAN model is close to algorithm utilizing evolutionary information (Supplementary Figures S8 and S9). These patterns demonstrate the efficacy of predictive assumption defined by separate training dataset. Taken together, the improvement of our composite models could be attributed to multiple incorporated properties, including the learning algorithm, the comprehensive aggregation of existing prediction scores as well as the different predictive assumptions defined by the training datasets.

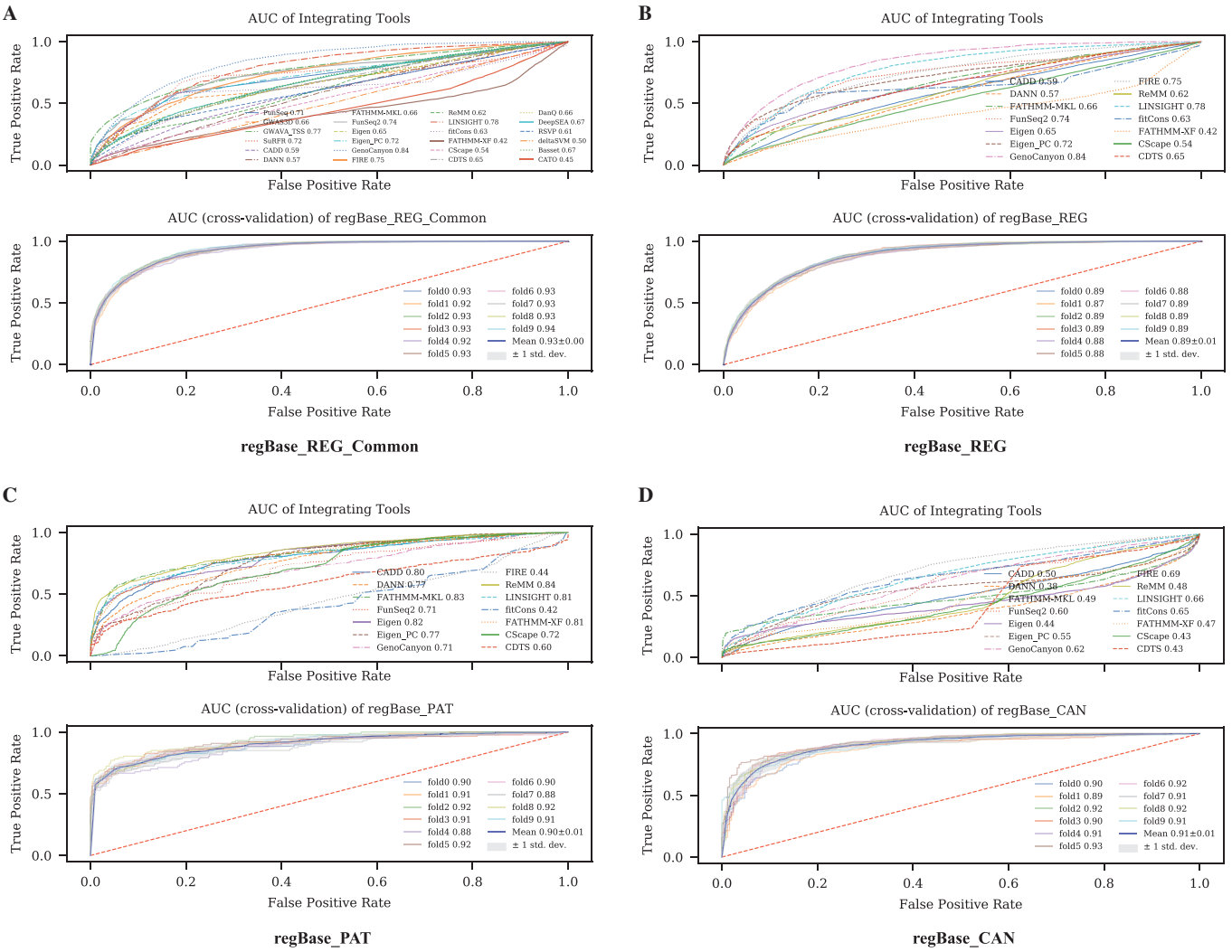

Figure 2. Receiver operating characteristic (ROC) curve and area under the receiver operating characteristics curve (AUC) for different prediction models using 10-fold cross-validation. (**A**) ROC and AUC of 23 integrated tools and 10-fold cross-validation result for regBase_REG_Common model. (**B**) ROC and AUC of 13 integrated tools and 10-fold cross-validation result for regBase_REG model. (**C**) ROC and AUC of 13 integrated tools and 10-fold cross-validation result for regBase_PAT model. (**D**) ROC and AUC of 13 integrated tools and 10-fold cross-validation result for regBase_CAN model.

Benchmarks on independent non-coding regulatory variant datasets

To systematically evaluate our four composite models, we constructed eight independent benchmark datasets across different functional categories of non-coding regulatory variants (Supplementary Table S7), including two fine-mapped eQTL datasets (Brown_eQTL (50), GTEx_eQTL (52)), one experimental validated eQTL dataset (MPRA_eQTL (44)), two disease-associated variants datasets (GWAS_5E-8, GWAS_1E-5 (53)), one somatic eQTL dataset (Somatic_eQTL (54)) and two pathogenic mutation dataset (Rare_Patho_SNV (22,55), ASD_denovo_SNV (56)). We also sampled corresponding control testing dataset and removed variants that appeared in our training datasets. These independent datasets were not used to train almost all of integrated algorithms in the regBase database, which could provide an unbiased opportunity to comprehensively compare our models with existing tools.

In general, our composite models can achieve an AUC score around 0.8–0.9 for most of the above testing sets. Among them, regBase_REG_Common model was the best one to predict fine-mapped eQTLs (AUC of 0.88 for Brown_eQTL, AUC of 0.89 for GTEx_eQTL) and GWAS disease-associated SNVs (AUC of 0.88 for GWAS_5E-8, AUC of 0.83 for GWAS_1E-5) (Figure 3A), while the performance regBase_REG is similar but falls slightly behind (Figure 3B). This is consistent with the cross-validation results in model training step. Interestingly, regBase_PAT model exhibited poor performance when predicting GWAS disease-associated variants. Compared with common germline variants that conferring hereditary disease predisposition, the pathogenic SNVs used to train regBase_PAT model are mostly rare variants to cause Mendelian disorders and obtain very distinct attributes. As

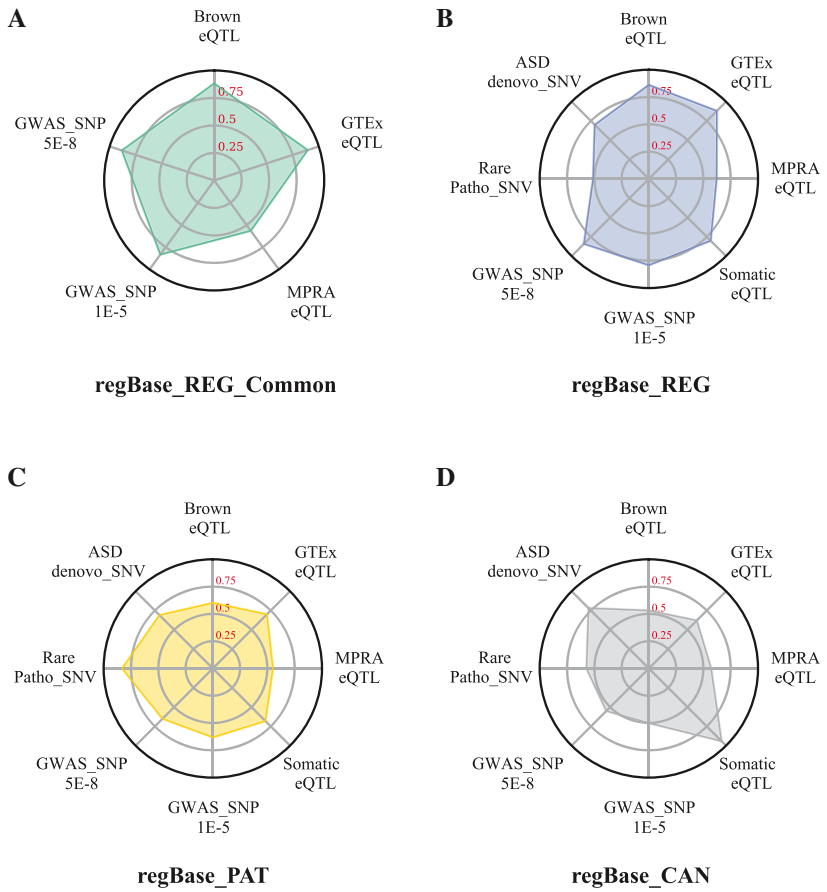

Figure 3. Area-under-curve scores distribution for eight independent benchmarks. (**A**) regBase_REG_Common model. (**B**) regBase_REG model. (**C**) regBase_PAT model. (**D**) regBase_CAN model. Brown_eQTL, 11 tissue/cell type-specific eQTLs fine-mapping data that was profiled by Brown and colleagues; GTEx_eQTL, 44 tissues-specific eQTLs within fine-mapped credible set from GTEx V6; MPRA_eQTL, significant expression modulating variants by MPRA in lymphoblastoid cell lines; GWAS_5E-8, GWAS disease-associated regulatory variants with *P*-value < 5E–8 from GWAS Catalog; GWAS_1E-5, GWAS disease-associated regulatory variants with *P*-value < 1E-5 from GWAS Catalog; Somatic_eQTL, recurrent somatic mutations within significant flanking intervals per somatic eGene; Rare_Patho_SNV, rare pathogenic regulatory variants for inherited diseases; ASD_denovo_SNV, *de novo* pathogenic regulatory mutations for autism spectrum disorder.

expected, regBase_PAT model outperformed other predictions (AUC of 0.83 for Rare_Patho_SNV) in discriminating rare pathogenic variants (Figure 3C). Regarding to the prediction of cancer relevant somatic eQTLs, regBase_CAN model received an AUC of 0.94 which largely outperformed other models (Figure 3D). In addition, regBase_CAN model also showed satisfactory performance (AUC of 0.78 for ASD_denovo_SNV) to predict pathogenic *de novo* mutations, further indicating the combination of individual classifiers could generate stronger learner using Gradient Tree Boosting strategy (Figure 3D). For predicting expression-modulating variants identified by MPRA, the best composite model regBase_REG got relatively smaller AUC of 0.62, implying the integration of existing tools may have limited ability to distinguish sequence effect of transcriptional regulatory elements regardless of their chromatin context.

To figure out whether the combined models are better than individual tools or not, we evaluated the performance of 23 regBase Common integrated scores on five common variants testing sets, and 13 regBase integrated scores three rare/*de novo*/somatic mutation datasets. Results showed that our composite models outperformed individual tools on most of evaluations. First, regBase_REG_Common model was top ranked for Brown_eQTL (Figure 4A and Supplementary Table S8), GTEx_eQTL (Figure 4B and Supplementary Table S9), GWAS_5E-8 (Figure 4C and Supplementary Table S10) and GWAS_1E-5 (Supplementary Figure S10A and Supplementary Table S11). It is worth noting that GenoCanyon, FIRE, LINSIGHT and Eigen_PC were well performed on predicting germline *cis*-eQTLs, while GenoCanyon, FunSeq2 and SuRFR were suitable to classify disease-associated regularity variants. In addition, regBase_PAT model preceded other predictions for Rare_Patho_SNV dataset, demonstrating its potential clinical significance to interpret rare regulatory variants causing inherited disease (Figure 4D and Supplementary Table S12). Third, regBase_CAN model was the best one for Somatic_eQTL dataset, with an AUC of 0.94 which greatly surpassed the second-best tool Eigen_PC (AUC of 0.86) (Figure 4E and Supplementary Table S13). regBase_CAN model also performed well with the highest AUC for ASD_denovo_SNV dataset, implying the shared

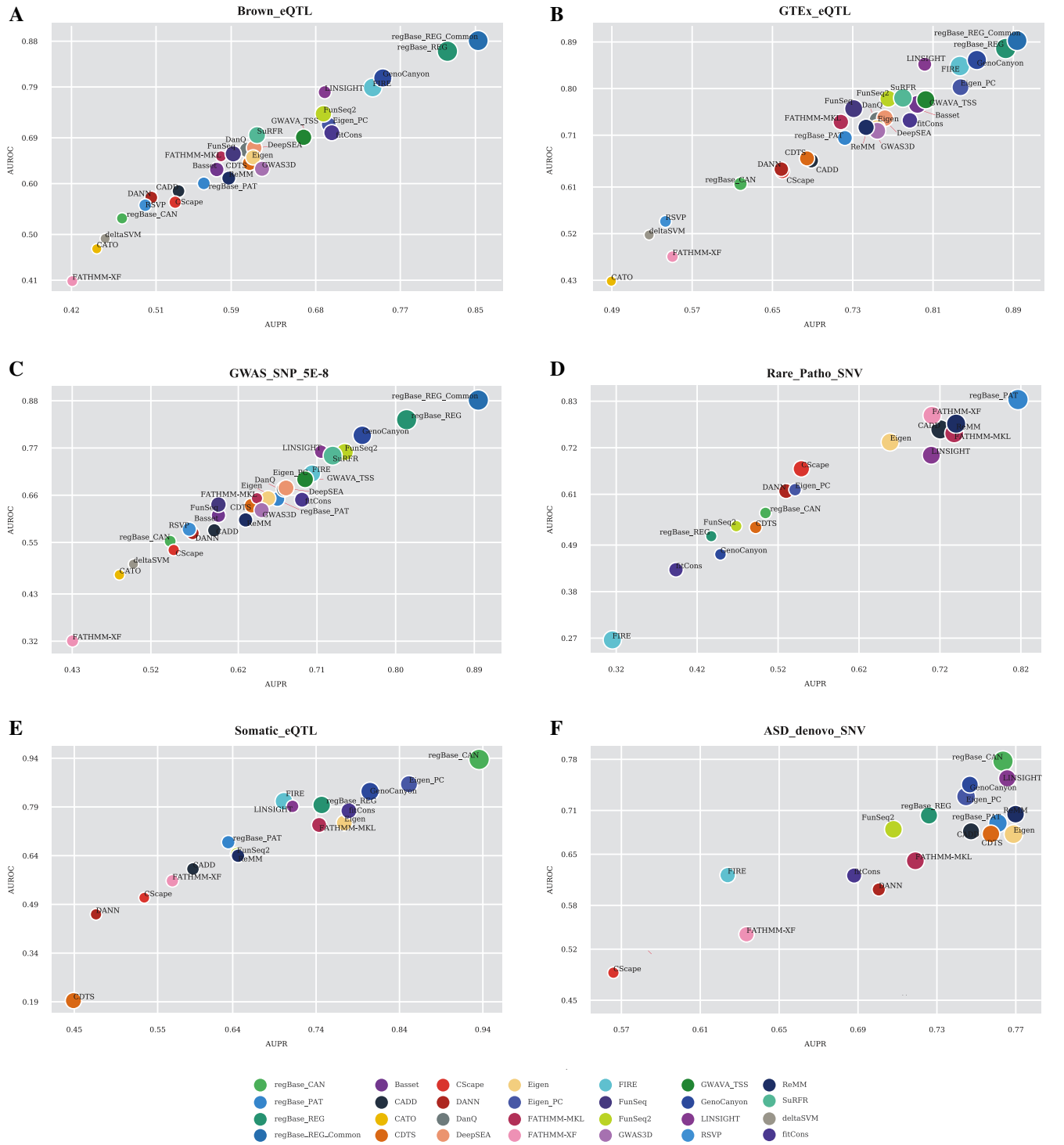

Figure 4. Evaluation result of individual prediction tools on six independent testing datasets. (**A**) Performance on Brown_eQTL dataset. (**B**) Performance on GTEx_eQTL dataset. (**C**) Performance on GWAS_5E-8 dataset. (**D**) Performance on Rare_Patho_SNV dataset. (**E**) Performance on Somatic_eQTL dataset. (**F**) Performance on ASD_denovo_SNV dataset. AUPR, area under the precision recal curve; AUROC, area under the receiver operating characteristics curve; bubble size is proportional to Pearson correlation coefficients between predicted and true labels for each evaluation.

regulatory properties between cancer driver somatic mutation and pathogenic de *novo* mutation (Figure 4F and Supplementary Table S14).

Moreover, when predicting effective MPRA alleles, tools learned by deep learning or unsupervised model, such as DeepSEA, GenoCanyon, Eigen_PC and Basset, obtained a higher AUC than our regBase_REG model (Supplementary Figure S10B and Supplementary Table S15), probably due to the fact that deep learning and unsupervised methods could capture unknown features that explain the *in vitro* activity of regulatory allele. Given the overall poor performance of existing tools and our composite models in predicting MPRA positive regulatory variants, we have retrained independent composite models, regBase_MPRA and regBase_MPRA_Common, using previously collected MPRA_eQTL dataset (44) to investigate whether improvement could be made. Comparing with existing methods, we did find slight improvements (∼3%) using cross-validation (Supplementary Figures S11 and S12). We also curated MPRA positive variants from other publications (56–58) and sampled strict matched controls, called MPRA_intergrated_SNV dataset. Using this independent test dataset, we found that our regBase_MPRA and regBase_MPRA_Common exhibited the best but still moderate performance to predict *in vitro* activity of regulatory allele (Supplementary Tables S16 and S17). This may suggest that accurate prediction of MPRA positive regulatory variants requires additional key features which are able to capture real context around assayed sequences.

We also evaluated the performance of our newly trained models with existing ensemble methods including IW-Scoring (12) and our previous PRVCS (11). We found that regBase_REG_Common model obtained superior capability in eQTL and GWAS regulatory variant benchmarks, except that PRVCS and IW-Scoring slightly outperformed regBase_REG model at MPRA_eQTL dataset. For pathogenic datasets, our composite models still largely outperformed other ensemble methods (Supplementary Figure S13 and Supplementary Table S18). Taken together, these independent evaluations further demonstrated the effectiveness of our composite models and illuminated that non-coding regulatory variants prediction results could be increasingly applicable in the future genetic studies.

regBase composite models facilitate the identification of causal non-coding regulatory variant from complex GWAS loci

Exploiting the true disease-causal variants is a challenging task in the GWAS study, especially for extremely high LD variants that locate in the non-coding genomic region. Statistical fine-mapping analysis usually ends with credible set of likely casual variants in which highly linked SNPs achieve similar posterior probabilities of causality, requiring further investigation of the true causal variants by other computational strategies, such as functional annotation (61). By visualizing regional PHRED-scaled score spectrum of composite models across 5p15.33 TERT region, we found several PHRED score peaks of regBase_REG, regBase_REG_Common and regBase_CAN generally colocalize with significant disease-associated variants identified by existing GWASs, especially in the TERT promoter region (Figure 5A and Supplementary Table S19). To evaluate the ability of our composite models for causal variant prioritization, we collected 22 unique SNPs in the 5p15.33 TERT region that confer risk of multiple cancers from ten GWAS fine-mapping results (Supplementary Table S20). Previous results showed there are many independent causal SNPs around the TERT genomic region, and many of them can alter promoter or enhancer activities (62). We revealed that our regBase_CAN and regBase_REG_Common models acquired relatively higher regional PHRED scores than other methods (tools with no >25% equal scores were selected) for collected fine-mapped SNPs (Figure 5B and Supplementary Table S21). Moreover, compared with relatively higher correlation among these 22 fine-mapped SNVs (Supplementary Figure S14), our top ranked variants (regional PHRED score > 10) of regBase_CAN or regBase_REG_Common showed very low LD with each other (Figure 5C), which indicates that our composite models could distinguish true signal from difficult credible set. For example, among all 22 prioritized fine-mapped SNPs by regBase_REG_Common model, rs2853669 obtained the largest PHRED score in the whole 5p15.33 TERT region (Figure 5C). This SNP was previously validated to disrupt TERT promoter and confer cancer risk by extensive functional experiments (63–65), further suggesting our composite model could efficiently narrow down the potentially causal variants for following functional validations.

regBase composite models discriminate casual regulatory alleles at base-wise resolution

To evaluate the ability of our composite models in distinguishing the true casual allele at base-wise level, we performed two independent comparisons using real and simulated datasets. First, recent studies of saturation mutagenesis could identify allele-specific effect for all possible sites of regulatory element (60,66). We selected a previously reported *ALDOB* (aldolase B, fructose-bisphosphate) enhancer which showed larger mutation effect in the saturation mutagenesis assay (60), and we compared whether our predicted scores are more correlated with the base-wise fold changes of experiment than scores from other single method. Since base-wise evaluation ideally requires non-missing and unique score at each site, we found that prediction scores of 13 regBase-incorporated tools for 259bp *ALDOB* enhancer overall showed high non-missing rate but some of them exhibited low uniqueness (Figure 6A). To ensure a valid base-wise comparison, we excluded tools with low score uniqueness (<75%) and performed correlation analysis between prediction scores and true fold changes of experiment. We showed that regBase_PAT model (Pearson correlation coefficients, $R = 0.4603$) outperformed all qualified prediction scores (Figure 6B) and other composite models (Supplementary Figure S15), which indicates the improved ability of our aggregated score in characterizing base-wise effect for regulatory element. Since *ALDOB* is a disease-causal gene of hereditary fructose intolerance (67), this result may also imply that the top-ranked regBase_PAT model could better distinguish pathogenic regulatory alleles than other methods.

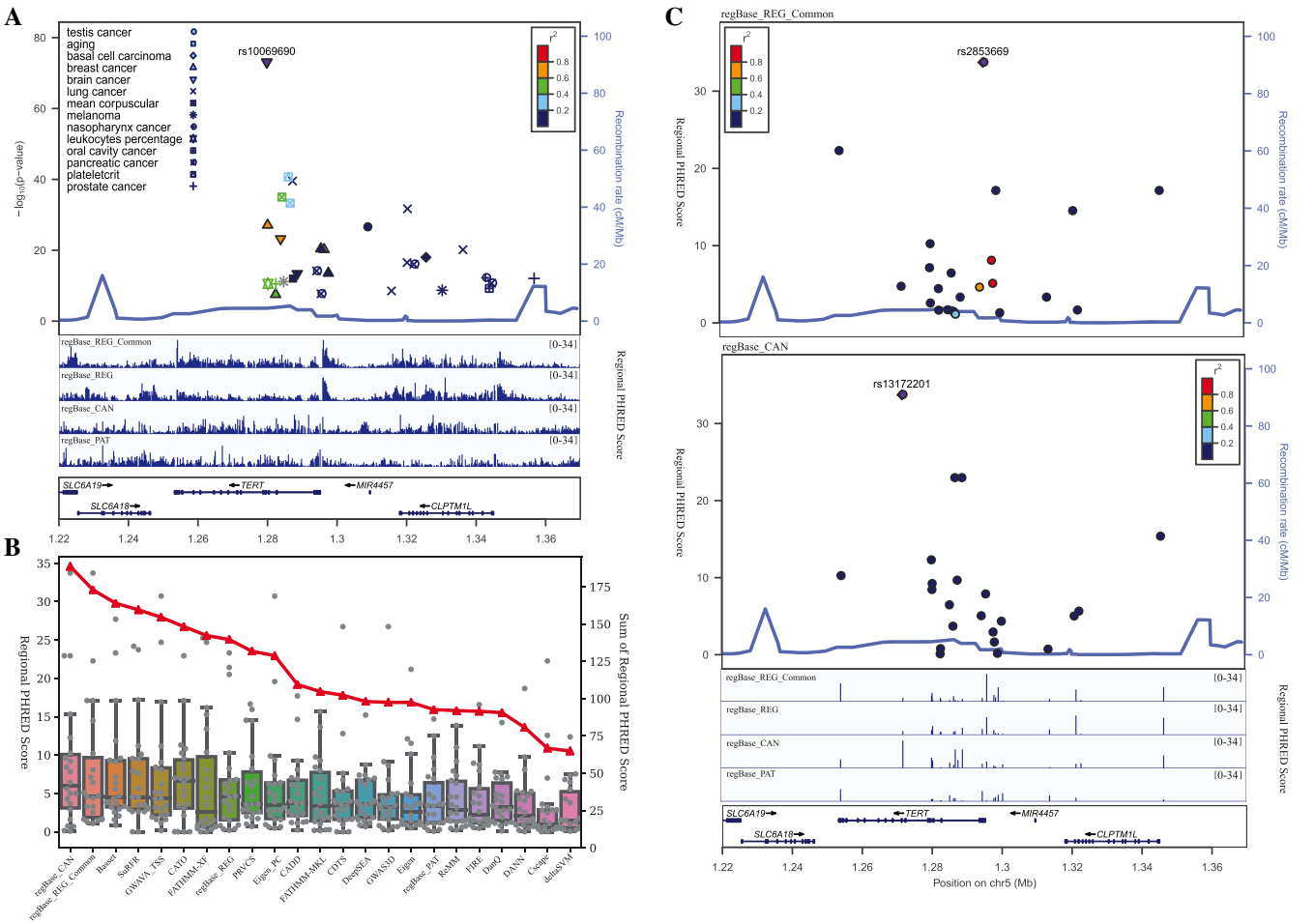

Figure 5. Non-coding regulatory variants prioritization at 5p15.33 TERT region. (**A**) GWAS significant SNPs and regional PHRED-scaled score distribution of our four composite models across 5p15.33 TERT region. LocusZoom plot is generated using the most significant SNP rs10069690 as lead and the EUR LD structure. (**B**) Comparison of regional PHRED scores among our composite models and all integrated methods for 22 fine-mapping SNPs at 5p15.33 TERT gene. Tools that obtain more than 25% equal scores in the evaluation are excluded. (**C**) LocusZoom plots for regional PHRED-scaled score of 22 fine-mapping SNPs. The top prioritized SNP rs2853669 in regBase_REG_Common model and the top prioritized SNP rs13172201 in regBase_CAN models are selected as leads.

Second, we collected a recently simulated 55 453 non-coding SNVs with pathogenic allele never observed in 57 diverse non-human placental mammals (typically evolutionarily forbidden alleles under purifying selection) and matched non-pathogenic derived alleles with frequencies of 5–15% in human (minimize potential influence by positive or balancing selection) at same position (13). Upon this simulated dataset, previous benchmark observed very low AUC of existing methods and concluded that biological usefulness of existing prediction scores for discriminating pathogenic alleles at single variant resolution is extremely limited (13). These inabilities could be attributed to several potential factors such as the false positives/negatives of simulated pathogenic/neutral alleles, the low uniqueness or limited discrimination of allelic prediction scores at same position, etc. As expected, majority of existing tools frequently predict the same score between simulated pathogenic and non-pathogenic alleles, and only six prediction tools show score difference for >50% sites (including our three composite models, Figure 6C). By prioritizing the distance of normalized prediction score at each position, we evaluated the capability of discriminating variant-level pathogenicity for six qualified models from a different angle. We found our regBase_PAT model achieves better degree of discrimination for top 1% prioritized variants, while regBase_CAN model works better for top 10% prioritized variants as a whole (Figure 6D), which reveals that our composite models may have higher discriminability in pathogenic allele detection at single variant resolution. Similar results were also observed when using additional simulated datasets by requiring that pathogenic alleles were sampled in different manners (Supplementary Figures S16 and S17).

DISCUSSION

Evolved methods had been developed to predict and prioritize functional non-coding regulatory variants, yet systematical integration of existing predicted scores for all possible substitutions of human SNV was largely deficient. Comparing with a commonly used lightweight resource dbNSFP

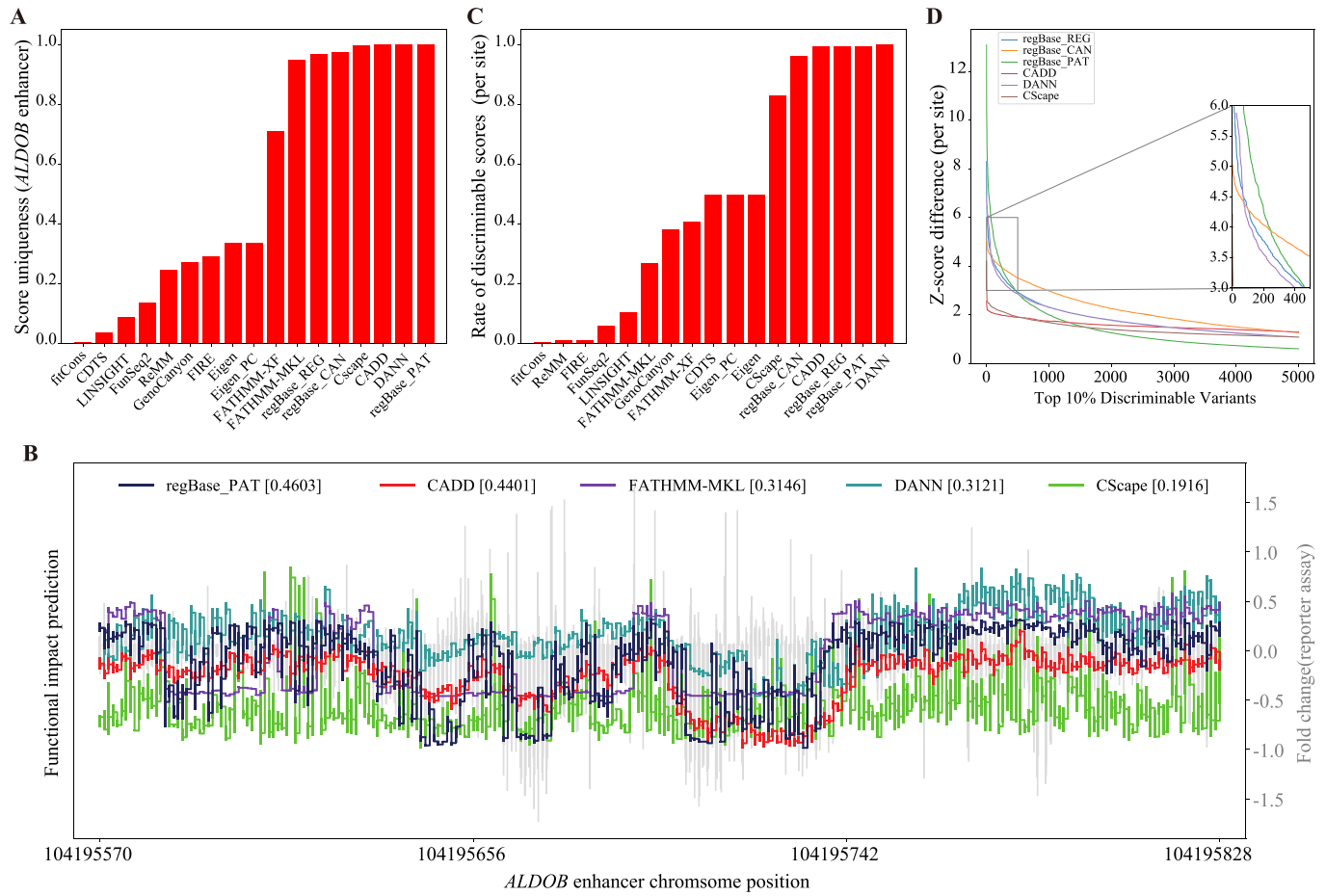

Figure 6. Causal regulatory alleles discrimination at base-wise resolution. (**A**) The uniqueness of prediction scores of 13 regBase-incorporated tools in the 259 bp *ALDOB* enhancer. (**B**) Prediction scores overlaid with expression fold changes (gray bars) for an *ALDOB* enhancer as determined with saturation mutagenesis assay. Pearson correlation values for this region are provided in parentheses for each method. (**C**) The proportion of discriminable scores among 13 regBase-incorporated tools for 55 453 simulated sites. (**D**) Degree of discrimination for pathogenic and non-pathogenic alleles of top prioritized variants among qualified prediction models.

on functional prediction and annotation for human non-synonymous and splice-site SNVs (68), we compile a comprehensive resource that includes 23 different tools to predict functional non-coding regulatory variants at the whole genome scale. To maximize the power and completeness for different types of non-coding regulatory variant prediction, we introduce three independent ensemble models to score functional, pathogenic or cancer driver regulatory variants respectively. We demonstrate that our composite strategies significantly increase the prediction accuracy and can greatly assist the casual non-coding regulatory variant discovery at base-wise resolution.

According to the benchmarks of several independent datasets, we found stable and reasonable performance of existing tools to predict variant regulatory potential regardless of its pathogenicity, such as predicting the probability of SNV to be a *cis*-eQTL. This merit could be attributed to the fact that current models are generally learned from annotation features that delineate regulatory signals around SNV locus, including chromatin accessibility, histone modifications and transcription factor binding. However, when evaluating the expression-modulating variants identified by *in vitro* reporter assay (60), no methods can achieve satisfactory performance. Since effective alleles in the MPRA are only weakly correlated with the associated eQTL effects (44,57), it may imply that surrounding sequence and local chromatin state could change the effect size of casual allele. In addition, recent CRISPR screening and GWAS fine mapping study have uncovered that some regulatory alleles locating in the unmarked regulatory elements are not associated with the conventional histone modifications or chromatin accessibility (69,70), which highlights the importance to exploit the missing but distinct prediction features. Besides, rational classification of pathogenic non-coding regulatory variant will extend the scopes of genetic diagnosis and precision medicine. Increasing studies have reported that pathogenic non-coding regulatory variant can influence the penetrance and causality of certain diseases (6), or alter the drug sensitivities (71,72). However, using ClinVar or COSMIC non-coding regulatory SNVs (not including splicing-altered SNVs) as golden standards (42,73), previous and our evaluations on pathogenic classification of regulatory variants showed limited performance (8,11). To this end, by leveraging the complementarity and unique-

ness of existing methods, we trained regBase_PAT and regBase_CAN models to score the probability of variants being pathogenic or cancer driver in the gene regulation, and found significant improvements in both cross-validation and independent benchmark. As the continual discoveries of non-coding disease-casual regulatory variants and more associated features, we believe that pathogenic prediction of non-coding regulatory variants will play a critical role in the clinical consensus interpretation of whole genome DNA sequence.

Highly context-dependent gene regulation can determine the cellular function of regulatory variants, and many recent methods are able to interpret regulatory variant in tissue/cell type-specific and disease-specific conditions (7,74). Since very few context-specific dataset could be used to benchmark the performance of tissue/cell type-specific predictions, researchers usually apply indirect solutions to evaluate the algorithms, such as the enrichment of tissue/cell type-specific epigenetic signals and *cis*-regulatory elements (75). Such imperfections and under calibrated performance could inhibit the broader applications of context-specific methods, especially for accurately predicting pathogenic regulatory variant on particular conditions. Despite the importance of systematic integration and evaluation of tissue/cell type-specific methods, regBase particularly aggregates and operates context-free prediction scores from existing tools. Our regBase aggregated scores together with three ensemble models provide a versatile tool that prioritizes organismal level non-coding regulatory variants in a context-free manner, greatly facilitating the interpretation of human non-coding genome in the era of precision medicine.

DATA AVAILABILITY

The regBase models are implemented in Python. Integrated datasets, source codes, collected training/testing sets, analysis scripts for the results of this manuscript are available at https://github.com/mulinlab/regBase.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China [31701143, 31871327 to M.J.L.]; Natural Science Foundation of Tianjin [18JCZDJC34700 to M.J.L.]; The Science & Technology Development Fund of Tianjin Education Commission for Higher Education [2018KJ082 to S.Z.]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

1. Gallagher,M.D. and Chen-Plotkin,A.S. (2018) The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, **102**, 717–730.
2. Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.*, **24**, R102–R110.
3. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
4. Weedon,M.N., Cebola,I., Patch,A.M., Flanagan,S.E., De Franco,E., Caswell,R., Rodriguez-Segui,S.A., Shaw-Smith,C., Cho,C.H., Allen,H.L. *et al.* (2014) Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.*, **46**, 61–64.
5. Li,X., Kim,Y., Tsang,E.K., Davis,J.R., Damani,F.N., Chiang,C., Hess,G.T., Zappala,Z., Strober,B.J., Scott,A.J. *et al.* (2017) The impact of rare variation on gene expression across tissues. *Nature*, **550**, 239–243.
6. Castel,S.E., Cervera,A., Mohammadi,P., Aguet,F., Reverter,F., Wolman,A., Guigo,R., Iossifov,I., Vasileva,A. and Lappalainen,T. (2018) Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.*, **50**, 1327–1334.
7. Rojano,E., Seoane,P., Ranea,J.A.G. and Perkins,J.R.(2018) Regulatory variants: from detection to predicting impact. *Brief. Bioinform*, doi:10.1093/bib/bby039.
8. Drubay,D., Gautheret,D. and Michiels,S. (2018) A benchmark study of scoring methods for non-coding mutations. *Bioinformatics*, **34**, 1635–1641.
9. Nishizaki,S.S. and Boyle,A.P. (2017) Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends Genet.*, **33**, 34–45.
10. Li,M.J., Yan,B., Sham,P.C. and Wang,J. (2015) Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief. Bioinform.*, **16**, 393–412.
11. Li,M.J., Pan,Z., Liu,Z., Wu,J., Wang,P., Zhu,Y., Xu,F., Xia,Z., Sham,P.C., Kocher,J.P. *et al.* (2016) Predicting regulatory variants with composite statistic. *Bioinformatics*, **32**, 2729–2736.
12. Wang,J., Dayem Ullah,A.Z. and Chelala,C. (2018) IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome. *Nucleic Acids Res.*, **46**, e47.
13. Liu,L., Sanderford,M.D., Patel,R., Chandrashekar,P., Gibson,G. and Kumar,S. (2019) Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.*, **10**, 330.
14. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
15. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**:D886–D894.
16. Ritchie,G.R., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
17. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
18. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.*, **47**, 276–283.
19. Kircher,M. and Shendure,J. (2015) Running spell-check to identify regulatory variants. *Nat. Genet.*, **47**, 853–855.
20. Maurano,M.T., Haugen,E., Sandstrom,R., Vierstra,J., Shafer,A., Kaul,R. and Stamatoyannopoulos,J.A. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat. Genet.*, **47**, 1393–1401.
21. Zhang,W., Bojorquez-Gomez,A., Velez,D.O., Xu,G., Sanchez,K.S., Shen,J.P., Chen,K., Licon,K., Melton,C., Olson,K.M. *et al.* (2018) A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.*, **50**, 613–620.
22. di Iulio,J., Bartha,I., Wong,E.H.M., Yu,H.C., Lavrenko,V., Yang,D., Jung,I., Hicks,M.A., Shah,N., Kirkness,E.F. *et al.* (2018) The human noncoding genome defined by genetic diversity. *Nat. Genet.*, **50**, 333–337.
23. Rogers,M.F., Shihab,H.A., Gaunt,T.R. and Campbell,C. (2017) CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.*, **7**, 11597.
24. Quang,D., Chen,Y. and Xie,X. (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, **31**, 761–763.

25. Ionita-Laza,I., McCallum,K., Xu,B. and Buxbaum,J.D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.*, **48**, 214–220.
26. Shihab,H.A., Rogers,M.F., Gough,J., Mort,M., Cooper,D.N., Day,I.N., Gaunt,T.R. and Campbell,C. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*, **31**, 1536–1543.
27. Rogers,M.F., Shihab,H.A., Mort,M., Cooper,D.N., Gaunt,T.R. and Campbell,C. (2018) FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, **34**, 511–513.
28. Ioannidis,N.M., Davis,J.R., DeGorter,M.K., Larson,N.B., McDonnell,S.K., French,A.J., Battle,A.J., Hastie,T.J., Thibodeau,S.N., Montgomery,S.B. *et al.* (2017) FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics*, **33**, 3895–3901.
29. Lu,Q., Hu,Y., Sun,J., Cheng,Y., Cheung,K.H. and Zhao,H. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.*, **5**, 10576.
30. Huang,Y.F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
31. Smedley,D., Schubach,M., Jacobsen,J.O.B., Kohler,S., Zemojtel,T., Spielmann,M., Jager,M., Hochheiser,H., Washington,N.L., McMurry,J.A. *et al.* (2016) A whole-genome analysis framework for effective identification of pathogenic regulatory variants in mendelian disease. *Am. J. Hum. Genet.*, **99**, 595–606.
32. Genomes Project, C., Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
33. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
34. Quang,D. and Xie,X. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**, e107.
35. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
36. Lee,D., Gorkin,D.U., Baker,M., Strober,B.J., Asoni,A.L., McCallion,A.S. and Beer,M.A. (2015) A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.*, **47**, 955–961.
37. Khurana,E., Fu,Y., Colonna,V., Mu,X.J., Kang,H.M., Lappalainen,T., Sboner,A., Lochovsky,L., Chen,J., Harmanci,A. *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, **342**, 1235587.
38. Peterson,T.A., Mort,M., Cooper,D.N., Radivojac,P., Kann,M.G. and Mooney,S.D. (2016) Regulatory single-nucleotide variant predictor increases predictive performance of functional regulatory variants. *Hum. Mutat.*, **37**, 1137–1143.
39. Ryan,N.M., Morris,S.W., Porteous,D.J., Taylor,M.S. and Evans,K.L. (2014) SuRFing the genomics wave: an R package for prioritising SNPs by functionality. *Genome Med.*, **6**, 79.
40. Li,H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
41. Stenson,P.D., Mort,M., Ball,E.V., Shaw,K., Phillips,A. and Cooper,D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.
42. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
43. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
44. Tewhey,R., Kotliar,D., Park,D.S., Liu,B., Winnicki,S., Reilly,S.K., Andersen,K.G., Mikkelsen,T.S., Lander,E.S., Schaffner,S.F. *et al.* (2018) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, **172**, 1132–1134.
45. Lesurf,R., Cotto,K.C., Wang,G., Griffith,M., Kasaian,K., Jones,S.J., Montgomery,S.B., Griffith,O.L. and Open Regulatory Annotation, C. (2016) ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.*, **44**, D126–D132.
46. Farh,K.K., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shoresh,N., Whitton,H., Ryan,R.J., Shishkin,A.A. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
47. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
48. Chen,T. and Guestrin,C. (2016) *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco, pp. 785–794.
49. Lundberg,S.M. and Lee,S.-I. (2017) A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, **30**, 4765–4774.
50. Brown,C.D., Mangravite,L.M. and Engelhardt,B.E. (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. *PLos Genet.*, **9**, e1003649.
51. Hormozdiari,F., Kostem,E., Kang,E.Y., Pasaniuc,B. and Eskin,E. (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.
52. Consortium, G.T.Laboratory, D.A.Coordinating Center -Analysis Working, G.Statistical Methods groups-Analysis Working, G.Enhancing, G.g.Fund, N.I.H.C.Nih/NciNih/NhgriNih/NimhNih/Nida2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
53. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
54. Calabrese,C., Davidson,N.R., Fonseca,N.A., He,Y., Kahles,A., Lehmann,K.-V., Liu,F., Shiraishi,Y., Soulette,C.M., Urban,L. *et al.* (2018) Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. bioRxiv doi: https://doi.org/10.1101/183889, 12 March 2018, preprint: not peer reviewed.
55. Caron,B., Luo,Y. and Rausell,A. (2019) NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.*, **20**, 32.
56. Zhou,J., Park,C.Y., Theesfeld,C.L., Wong,A.K., Yuan,Y., Scheckel,C., Fak,J.J., Funk,J., Yao,K., Tajima,Y. *et al.* (2019) Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, **51**, 973–980.
57. Ulirsch,J.C., Nandakumar,S.K., Wang,L., Giani,F.C., Zhang,X., Rogov,P., Melnikov,A., McDonel,P., Do,R., Mikkelsen,T.S. *et al.* (2016) Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, **165**, 1530–1545.
58. Madan,N., Ghazi,A., Kong,X., Chen,E., Shaw,C. and Edelstein,L. (2019) Identification of functional variants for platelet CD36 expression by Massively Parallel Reporter Assay. bioRxiv doi: https://doi.org/10.1101/550871 , 15 February 2019, preprint: not peer reviewed.
59. Pruim,R.J., Welch,R.P., Sanna,S., Teslovich,T.M., Chines,P.S., Gliedt,T.P., Boehnke,M., Abecasis,G.R. and Willer,C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.
60. Patwardhan,R.P., Hiatt,J.B., Witten,D.M., Kim,M.J., Smith,R.P., May,D., Lee,C., Andrie,J.M., Lee,S.I., Cooper,G.M. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, **30**, 265–270.
61. Huang,D., Yi,X., Zhang,S., Zheng,Z., Wang,P., Xuan,C., Sham,P.C., Wang,J. and Li,M.J. (2018) GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. *Nucleic Acids Res.*, **46**, W114–W120.
62. Bell,R.J., Rube,H.T., Xavier-Magalhaes,A., Costa,B.M., Mancini,A., Song,J.S. and Costello,J.F. (2016) Understanding TERT promoter mutations: a common path to immortality. *Mol. Cancer Res.*, **14**, 315–323.

63. Rachakonda,P.S., Hosen,I., de Verdier,P.J., Fallah,M., Heidenreich,B., Ryk,C., Wiklund,N.P., Steineck,G., Schadendorf,D., Hemminki,K. *et al.* (2013) TERT promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17426–17431.
64. Spiegl-Kreinecker,S., Lotsch,D., Ghanim,B., Pirker,C., Mohr,T., Laaber,M., Weis,S., Olschowski,A., Webersinke,G., Pichler,J. *et al.* (2015) Prognostic quality of activating TERT promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis. *Neuro Oncol.*, **17**, 1231–1240.
65. Helbig,S., Wockner,L., Bouendeu,A., Hille-Betz,U., McCue,K., French,J.D., Edwards,S.L., Pickett,H.A., Reddel,R.R., Chenevix-Trench,G. *et al.* (2017) Functional dissection of breast cancer risk-associated TERT promoter variants. *Oncotarget*, **8**, 67203–67217.
66. Patwardhan,R.P., Lee,C., Litvin,O., Young,D.L., Pe'er,D. and Shendure,J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, **27**, 1173–1175.
67. Santer,R., Rischewski,J., von Weihe,M., Niederhaus,M., Schneppenheim,S., Baerlocher,K., Kohlschutter,A., Muntau,A., Posselt,H.G., Steinmann,B. *et al.* (2005) The spectrum of aldolase B (ALDOB) mutations and the prevalence of hereditary fructose intolerance in Central Europe. *Hum. Mutat.*, **25**, 594.
68. Liu,X., Wu,C., Li,C. and Boerwinkle,E. (2016) dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.*, **37**, 235–241.
69. Rajagopal,N., Srinivasan,S., Kooshesh,K., Guo,Y., Edwards,M.D., Banerjee,B., Syed,T., Emons,B.J., Gifford,D.K. and Sherwood,R.I. (2016) High-throughput mapping of regulatory DNA. *Nat. Biotechnol.*, **34**, 167–174.
70. Huang,H., Fang,M., Jostins,L., Umicevic Mirkov,M., Boucher,G., Anderson,C.A., Andersen,V., Cleynen,I., Cortes,A., Crins,F. *et al.* (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, **547**, 173–178.
71. Soccio,R.E., Chen,E.R., Rajapurkar,S.R., Safabakhsh,P., Marinis,J.M., Dispirito,J.R., Emmett,M.J., Briggs,E.R., Fang,B., Everett,L.J. *et al.* (2015) Genetic variation determines PPARgamma function and anti-diabetic drug response in vivo. *Cell*, **162**, 33–44.
72. Li,M.J., Yao,H., Huang,D., Liu,H., Liu,Z., Xu,H., Qin,Y., Prinz,J., Xia,W., Wang,P. *et al.* (2017) mTCTScan: a comprehensive platform for annotation and prioritization of mutations affecting drug sensitivity in cancers. *Nucleic Acids Res.*, **45**, W215–W221.
73. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
74. Li,M.J., Li,M., Liu,Z., Yan,B., Pan,Z., Huang,D., Liang,Q., Ying,D., Xu,F., Yao,H. *et al.* (2017) cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol.*, **18**, 52.
75. Roadmap Epigenomics, C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.