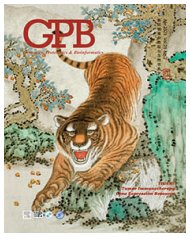
Genomics Proteomics Bioinformatics 21 (2023) 385–395

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



METHOD

RegVar: Tissue-specific Prioritization of Non-coding Regulatory Variants

Hao Lu, Luyu Ma, Cheng Quan, Lei Li, Yiming Lu *, Gangqiao Zhou *, Chenggang Zhang *

Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Beijing 100850, China

Received 11 January 2021; revised 11 June 2021; accepted 27 September 2021
Available online 29 December 2021

Handled by Leng Han

KEYWORDS

Non-coding variant;
Variant prioritization;
Expression regulation;
Expression quantitative trait locus;
Deep neural network

Abstract Non-coding genomic variants constitute the majority of trait-associated genome variations; however, the identification of functional non-coding variants is still a challenge in human genetics, and a method for systematically assessing the impact of regulatory variants on gene expression and linking these regulatory variants to potential target genes is still lacking. Here, we introduce a deep neural network (DNN)-based computational framework, RegVar, which can accurately predict the tissue-specific impact of non-coding regulatory variants on target genes. We show that by robustly learning the genomic characteristics of massive variant–gene expression associations in a variety of human tissues, RegVar vastly surpasses all current non-coding variant prioritization methods in predicting regulatory variants under different circumstances. The unique features of RegVar make it an excellent framework for assessing the regulatory impact of any variant on its putative target genes in a variety of tissues. RegVar is available as a web server at https://regvar.omic.tech/.

Introduction

Trait-associated genetic variants usually lie within non-coding genomic regions [1,2], and the interpretation of functional non-coding variants is crucial for revealing the underlying genetic architecture and molecular mechanism of complex traits and diseases. Several methods such as CADD [3], GWAVA [4], DeepSEA [5], and LINSIGHT [6] have been developed to discriminate pathogenic variants from nonpathogenic ones using genomic sequences, functional annotations, and evolutionary features. A common feature of these methods is that they focus on identifying rare pathogenic variants, which have been thought to have a stronger impact on human traits and diseases than common variants [7]. However, emerging evidence suggests that the major portion of heritability for complex traits is likely to be explained by a substantial number of common regulatory variants with small additive effect sizes, in combination with a relatively smaller contribution from rare variants of moderate effect sizes [8–10]. Thus, a model that can distinguish both common and rare regulatory variants will

* Corresponding authors.
E-mail: luym@bmi.ac.cn (Lu Y), zhougq@chgb.org.cn (Zhou G), zhangcg@bmi.ac.cn (Zhang C).

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2021.08.011

provide new perspectives on the regulatory basis of complex traits.

Current pathogenic variant prioritization models are not suitable for identifying regulatory variants. A recent survey of existing methods for prioritizing non-coding variants showed that, although they achieved high precision in identifying pathogenic variants under certain circumstances, their performance in identifying regulatory variants was very poor [11], because the prioritization of regulatory variants is an even greater challenge than the prioritization of pathogenic variants. First, regulatory variants generally have a weaker impact on gene expression than pathogenic variants, so it is more difficult to discriminate the regulatory variants from background, especially from adjacent nonfunctional variants sharing similar epigenetic marks. Second, it is challenging to link regulatory variants to their target genes, which can be located far away from their regulator. Third, it is a challenge to establish tissue- or cell type-specific models that can predict the regulatory impact of variants under different biological conditions. A number of methods have been proposed to predict the effects of regulatory variants in recent years [12–14]. However, these methods have limitations in their application. For example, ExPecto relies on epigenetic marks at gene promoters to monitor the regulatory impact of variants on gene expression and thus could only assess promoter-proximal variants [12]. TIVAN connects various genomic features to expression quantitative trait loci (eQTLs) to estimate the regulatory probability of a variant, but TIVAN was trained with promoter-proximal variants, which may introduce potential biases when applied to genome-wide variant prioritization [13]. Considering that the vast majority of regulatory variants are located far from the transcription start sites (TSSs) of target genes [15], a method that can robustly predict genome-wide regulatory variants as well as their potential target genes is urgently needed.

Here, we introduce a deep neural network (DNN)-based approach, RegVar, for the genome-wide assessment of the regulatory impact of non-coding variants on gene expression. RegVar has several key features: (1) it can predict both common and rare regulatory variants by learning their genomic characteristics from massive variant–gene associations in an unbiased manner; (2) it predicts not only regulatory variants but also their target genes by jointly learning the genomic patterns of both variants and genes and the chromatin interactions between them; (3) it predicts the tissue-specific effects of variants by training models in multiple tissues with respective genomic patterns; and (4) it can achieve excellent prediction accuracy by utilizing large training sets and deep learning algorithms. We show that RegVar outperforms existing prioritization methods in identifying regulatory variants and non-coding pathogenic variants from different backgrounds in various tissues.

Method

Datasets

To construct the positive datasets, significant eQTL variant (eVariant)–eQTL gene (eGene) associations in 17 human tissues that were also incorporated in the Roadmap Epigenomics Projects [2] were obtained from the Genotype-Tissue Expression (GTEx) project (v7 release) [16] (**Figure 1**; Table S1). Single-nucleotide variant (SNV)–eGene associations were selected and further filtered by removing eVariants not marked by DNase I hypersensitive site (DHS) annotations, which were demonstrated to be a key epigenetic marker of causal variants [16]. Associations in sex chromosomes were also removed. For tissues in which the number of significant associations exceeded 100,000 (esophagus mucosa, lung, skeletal muscle, and whole blood), we randomly selected 100,000 associations, as we found that a larger size did not improve model performance (Figure S1). The final number of positive associations for each tissue is shown in Table S1. For negative datasets, four datasets were constructed: (1) *random-variant* set of shuffled SNV–Gene pairs where eVariants were replaced by random SNVs located $\leq$ 1 Mb from the eGene TSS; (2) *mirrored-variant* set of shuffled pairs where eVariants were replaced by random SNVs located at similar distance (error $\leq$ 1 kb) but the opposite side of the eGene TSS; (3) *neighboring-variant* set of shuffled pairs where eVariants were replaced by random SNVs located adjacent ($\leq$ 1 kb) to the positive ones; and (4) *random-gene* set of shuffled SNV–Gene pairs where eGenes were replaced by gene TSSs located $\leq$ 1 Mb of the eVariants. We selected a maximum distance of 1 Mb between SNVs and TSSs in datasets (1) and (4) because all positive SNV–TSS pairs had a distance of less than 1 Mb (Figure S2). Variants in negative datasets were selected from the dbSNP build 146 after removing the shared variants between GTEx and dbSNP datasets.

Since eVariants are biased toward high-frequency variants (Figure S3), to ensure that our results were not influenced by the differences in minor allele frequency (MAF) between the positive variants and negative controls, we defined additional sets of MAF-matched negative controls for the GTEx liver dataset with the same strategy as described above.

Annotation profiles

We used three major categories of genomic profiles, including sequential, epigenetic, and evolutionary profiles (Table S2), to annotate our datasets using a customized pipeline.

Sequential profiles

Sequential profiles consisted of 2-mer prefix and postfix and local 5-mer GC content of SNV and TSS, SNV-caused transcription factor binding site (TFBS) affinity changes, genomic distance between SNV and TSS, and the orientations of SNV and TSS.

To calculate TFBS affinity changes caused by variants, we obtained the position frequency matrices of 602 transcription factor (TFs) from the TRANSFAC [17] (523 TFs) and JASPAR [18] (79 TFs) databases. TFM-Scan [19] was used to locate putative TFBS motifs by scanning genomic DNA both forward and backward using these position frequency matrices. A stringent threshold of $P < 4.5\text{E}{-5}$ was used to determine significant motifs. Variants located within these motifs were determined using BEDTools [20]. The TFBS affinity was calculated as previously described [21]. Specifically, the corrected probability of observing a given nucleotide in a specific locus was calculated as follows:

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,T,C,G\}} s(b')} \quad (1)$$

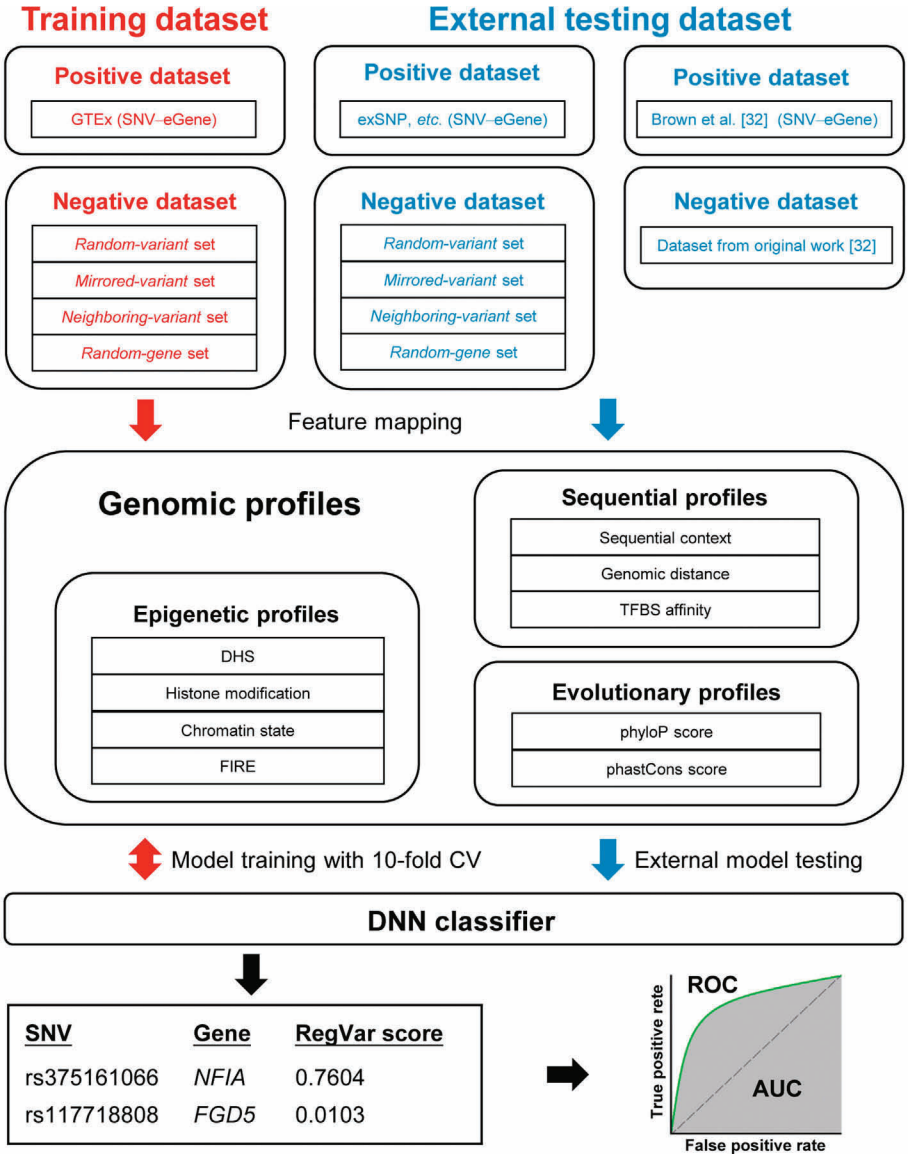


Figure 1 A flowchart showing the workflow of RegVar

GTEx, Genotype-Tissue Expression; SNV, single-nucleotide variant; eGene, expression quantitative trait locus gene; CV, cross-validation; DNN, deep neural network; TFBS, transcription factor binding site; FIRE, frequently interacting region; DHS, DNase I hypersensitive site; ROC, receiver operating characteristic; AUC, area under the ROC curve.

where b represents one specific base among A, T, C, and G, i is the index of the site, $f_{b,i}$ is the count of base b at site i, N is the sum of the counts of four bases, and $s(b)$ is the pseudocount function. Here, we assumed $s(b)$ to be 1/4 for each of the four bases, then:

$$\sum_{b' \in \{A,T,C,G\}} s(b') = 1 \tag{2}$$

Hence, the corresponding position weight matrix (PWM) can be constructed as:

$$W(b,i) = \lg \frac{p(b,i)}{p(b)} \tag{3}$$

where $p(b)$ is the background probability of base b (assumed to be 1/4 for each of the four bases). The TFBS affinity was calculated as follows:

$$Affinity = \sum_{i=1}^{w} W(b,i) \tag{4}$$

where w is the width of a PWM. We then calculated the average affinity change between reference and alteration alleles as follows:

$$\Delta Affinity = \frac{Affinity_A - Affinity_R}{w} \tag{5}$$

where $Affinity_R$ and $Affinity_A$ are evaluated binding affinities with the reference and alteration alleles, respectively. Variants located within two or more TFBS motifs were assigned with the $\Delta Affinity$ score with the maximum absolute value among all $\Delta Affinity$ scores of the affected motifs, and variants not located at any TFBS motif were assigned a $\Delta Affinity$ score of 0.

Epigenetic profiles

Epigenetic profiles consisted of 31 histone modifications from the Roadmap Epigenomics Project [2], 25 chromatin states produced by ChromHMM [22], and frequently interacting region (FIRE) annotations from the Hi-C study [23].

Evolutionary profiles

Evolutionary profiles consisted of vertebrate, placental mammal, and primate phyloP [24] and phastCons [25] scores based on 46-way whole-genome alignment and vertebrate phyloP and phastCons scores based on 100-way whole-genome alignment.

All annotations were expressed in genomic coordinates for the GRCh37/hg19 assembly of the human genome. Boolean variables were used to indicate whether SNVs or TSSs overlapped with chromatin marks (1) or not (0). For categorical annotations, all *n*-level categorical values were first encoded to binary values and then converted to several individual Boolean flags. For continuous annotations, feature values were scaled to the range of [0, 1]. More exactly, distance to TSS was scaled by:

$$distance' = \min(1, \lg(\mathrm{abs}(distance) + 1)/6) \quad (6)$$

phyloP scores were scaled by:

$$phyloP' = \min(1, \mathrm{abs}(phyloP)/5) \quad (7)$$

and $\Delta Affinity$ scores were scaled by:

$$\Delta Affinity' = \min(1, \mathrm{abs}(\Delta Affinity)) \quad (8)$$

Model design and training

We built a DNN-based classifier to model our dataset. The basic model in RegVar is a fully connected neural network, in which each neuron in a layer receives inputs from all outputs of the previous layer, except that the first layer receives inputs from the original data matrix. Each layer in the network executes a linear transformation of the corresponding inputs to integrate information from the previous layer, followed by a nonlinear transformation (namely, the activation function) to rectify the linear result. Here, we employed three fully connected layers with 500, 200, and 60 units, respectively, and the most commonly used rectified linear unit (ReLU) function as the activation function. Specifically, one fully connected layer computes the following:

$$output = \mathrm{ReLU}(WX + b) \quad (9)$$

where X is the input, W is the weight matrix, b is the bias, and ReLU represents the rectified linear function:

$$\mathrm{ReLU}(x) = \max(0, x) \quad (10)$$

The layer following the third fully connected layer is the final output layer to make predictions about being a regulatory or nonregulatory variant on the specific gene, with scaled probability ranging from 0 to 1 using:

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

To train the model, we selected the cross-entropy loss function as the objective function, which is defined as follows:

$$objective = -\frac{1}{N}\sum_{i=1}^{N}(Y_i * \log(f(X_i)) + (1 - Y_i) * \log(1 - f(X_i))) \quad (12)$$

where N is the number of samples in the training set, and i is the index of each sample. Y_i and X_i represent the 0/1 label and the input features for sample i, respectively, and $f(X_i)$ represents the predicted probability output from the DNN model.

We conducted an optimal search of hyperparameters, including the learning rate and dropout proportion. Learning rates were set at 0.001, 0.005, and 0.01; dropout proportions were set at 0, 0.3, and 0.5. We selected the combinations of learning rates and dropout proportions that achieved the highest prediction area under the receiver operating characteristic (ROC) curve (AUC) in each of the four models (Tables S3–S6).

All training programs were written in Python language using a DNN implementation from the TensorFlow library.

Model comparison

We used the average ROC curves computed from 10-fold cross-validation (CV) to evaluate model performance. Specifically, each dataset comprising the positive set and its negative counterpart was randomly split into a training set and a testing set in a 9:1 ratio; the RegVar model was trained on the training set and evaluated on the testing set. This process was repeated 10 times for each dataset, with an independent sample split procedure each time.

Predictions of CADD (v1.3) [3], GWAVA (v1.0) [4], DeepSEA [5], LINSIGHT [6], ExPecto [12], and TIVAN [13], together with two ensemble methods, IW-Scoring [26] and regBase [27], were used for model performance comparison in liver, hippocampus, and whole blood datasets. The *random-variant*, *mirrored-variant*, and *neighboring-variant* datasets were used for the evaluation, and the *random-gene* dataset was excluded because the existing methods did not predict potentially affected genes. In addition, ExPecto was excluded from the *random-variant* and *random-gene* model evaluation, because it focused on promoter-proximal variants, thus resulting in too few samples for the evaluation. For CADD, DeepSEA, and IW-Scoring, we ran the analysis using the corresponding online web services; for GWAVA, LINSIGHT, ExPecto, TIVAN, and regBase, we downloaded the precomputed scores from the corresponding source websites.

Model external evaluation

We downloaded liver eQTLs from the exSNP website [28], hippocampus eQTLs from Schulz [29] and Ramasamy [30] eQTL studies, and blood eQTLs from Westra [31] eQTL meta-analysis to evaluate the performance of trained models. We identified all SNV–TSS pairs and removed those overlapping with liver, hippocampus, and whole blood eQTLs in the GTEx dataset. For negative controls, all SNVs in the external positive datasets were removed from dbSNP build 146, and then four negative datasets were constructed, as described above in model training, for each of the three independent positive datasets. Additionally, the negative samples overlapping with the control sets used in model training were further removed to avoid any valid set contamination. Then, we annotated each

sample set with classifiers trained on GTEx eQTLs in the corresponding tissue and compared classification results with ROC curves for the first three sets with existing methods mentioned above.

In addition, we also performed model evaluation on the liver and blood eQTL data from the Brown eQTL dataset [32], which has been used for model evaluation by both regBase and TIVAN. We downloaded the original positive and negative sets for the Brown eQTL dataset from the regBase website, and compared the performance of different methods on the datasets of liver and blood tissues (all testing datasets are summarized in Table S7; see File S1 for more details about data processing).

Model evaluation on imbalanced datasets

To evaluate the impact of the ratio between positive and negative datasets on model training, we first assessed different ratios, including 1:1, 1:2, 1:3, 1:5, and 1:10, in the GTEx dataset. We found that there was no significant difference in performance among models trained with the different ratios (Figure S4). Thus, we selected a ratio of 1:1 between the positive and negative datasets to efficiently train the models. To demonstrate its application in imbalanced datasets, we also constructed a series of external testing datasets with positive: negative ratios of 1:1, 1:2, 1:3, 1:5, and 1:10 using liver eQTLs from the exSNP database. We showed that RegVar trained on GTEx dataset at the 1:1 ratio exhibited robust performance across these datasets with different ratios (Figure S5).

RegVar score distribution

For each tissue, we trained an integrated RegVar model by pooling four negative datasets to take all conditions together. Seventeen integrated RegVar models were applied to annotate all possible SNV–eGene pairs on chromosome 22. For each SNV on chromosome 22, we obtained TSSs of all genes located within 1 Mb of the variant locus and combined the variant with each of these TSSs as a possible eQTL pair. After mapping with all kinds of features, 65,844,726 sample pairs of 1,039,985 different SNVs remained and were annotated with integrated RegVar models in 17 tissue types. For each variant, the maximum annotated score of all its possible eQTL pairs was set as the final RegVar score of the variant in each tissue. We next explored the distribution patterns of RegVar scores of all these variants.

Tissue-shared/tissue-specific regulatory variants

Stratified random sampling was performed to select 100,000 SNVs from 22 autosomes, and TSSs located within 1 Mb of each variant locus were identified and combined with the variant as a possible eQTL pair. After mapping with the corresponding features, 3,703,900 sample pairs remained. RegVar scores were obtained in 17 tissue types and then converted to percentiles based on the corresponding merged training sets to make results comparable across different tissues. For a particular variant, the sample pair with the maximum percentile among all its possible eQTL pairs and across 17 integrated models was set to be the final sample pair. We obtained 17 tissue-specific percentiles of all final sample pairs to form a percentile matrix. Then, *K*-means clustering, implemented by the *kmeans* function in R language, was applied to the matrix to obtain tissue-specific and tissue-shared regulatory variant clusters. Four tissue-specific epigenetic features, namely, DHS, H3K4me1, H3K4me3, and H3K27ac, were used to annotate these tissue-specific and tissue-shared regulatory variants.

Results

RegVar shows better performance in prioritization of regulatory variants

To explore the influence of the DHS filter on the prediction capability of RegVar, in the liver dataset we first compared the performance of models built from the positive datasets: (1) with all features as predictors; (2) with DHS peaks alone as a predictor; (3) with the DHS peaks as the filter; and (4) with the assay for transposase-accessible chromatin using sequencing (ATAC-seq) peaks as the filter. For examination of potential bias from the negative datasets, we also constructed models built from both positive and negative datasets with the DHS peaks as the filter (File S1). We found that models built from the positive dataset with the DHS filter showed the most robust performance in discriminating regulatory variants from different backgrounds (Figure S6). We then utilized DHS filter-based RegVar to predict the tissue-specific effects of genomic variants on gene expression in 17 human tissues. The averaged ROC curves across 17 tissues showed that RegVar predicted regulatory variants and their target genes with averaged AUCs of 0.965, 0.917, 0.693, and 0.929 for the four training datasets, respectively (**Figure 2**; Table S1). This result demonstrates that RegVar could reliably discriminate positive regulatory variants from different negative backgrounds. We then evaluated the performance of existing methods CADD [3], GWAVA [4], DeepSEA [5], LINSIGHT [6], ExPecto [12], TIVAN [13], IW-Scoring [26], and regBase [27] on the same tasks. For CADD, we used C-scores. For GWAVA, we used pathogenic scores with the corresponding control standards (namely, *unmatched*, *TSS*, and *region*). For DeepSEA, we used eQTL-probability scores. For IW-Scoring, we used integrative scores without fitCons. For regBase, we used regBase_Common prediction scores. For the three tested tissues, liver, hippocampus, and whole blood, we found that only GWAVA, LINSIGHT, and IW-Scoring could make valid predictions with AUCs of 0.668–0.764 for *random-variant* set and 0.573–0.677 for *mirrored-variant* set, yet still much lower than RegVar (0.957–0.969 for *random-variant* set and 0.884–0.945 for *mirrored-variant* set), while the other five existing methods failed to show significant power in distinguishing regulatory variants (**Figure 3**A, Figures S7 and S8). For the *neighboring-variant* set, which is more challenging, none of the existing methods made valid predictions compared to RegVar with an AUC of 0.694–0.700. Precision-recall (PR) curves also showed that RegVar had a lower type I error than the other methods (Figures S9–S11). We also constructed MAF-controlled datasets to assess the impact of MAF on model training. After controlling for MAF, RegVar showed comparable prediction capabilities in discriminating eQTLs from MAF-matched benign variants, with only a slight decrease in the *neighboring-variant* set, as demonstrated in liver samples, and the other eight methods still showed low prediction capabilities (Figure S12).

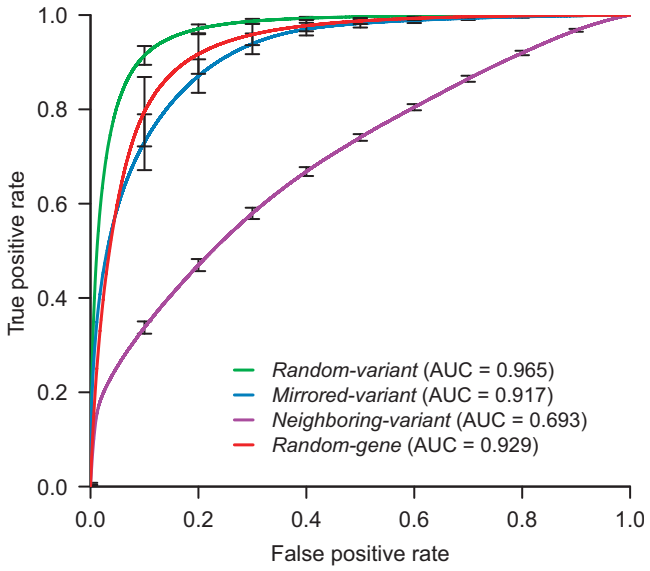


Figure 2 Average ROC curves for 10-fold CV experiments of RegVar

For each of the four training sets, the ROC curves were averaged across 17 human tissues. Error bars represent the standard deviation averaged over tissues.

To further confirm the results, we curated another three publicly available eQTL datasets of liver, hippocampus, and whole blood from the exSNP website [28], Schulz [29] and Ramasamy [30] eQTL studies, and Westra [31] eQTL meta-analysis, respectively, as independent testing sets. We found that RegVar models trained with GTEx datasets achieved almost equally accurate predictions in the three independent testing sets, while all other methods still did not show any obvious predictive powers in the independent datasets (Figure 3B, Figures S7–S11). In addition, we also evaluated the performance of different methods on the Brown eQTL data in liver and blood, which have been used as testing data for regBase and TIVAN. The results showed that in both tissues, RegVar trained on GTEx datasets showed comparable performance (AUC = 0.858 and 0.901 for liver and blood sets, respectively) with regBase (AUC = 0.883 and 0.890, equal to those reported in the study of regBase [27]). When trained on the Brown eQTL data, RegVar achieved even higher AUCs (AUC = 0.952 and 0.945 for liver and blood sets, respectively) compared with other methods (Figures S13 and S14). Altogether, these results demonstrate the outstanding performance of RegVar in predicting the regulatory impact of non-coding variants.

To investigate the robustness of RegVar on different settings of negative data sampling in external evaluation, we constructed negative datasets for the exSNP testing set by randomly selecting variants at wider genome regions, including: (1) *random-variant* set comprising random SNVs located $\leq$ 2 or 5 Mb from the eGene TSS; (2) *mirrored-variant* set comprising random SNVs with a distance error $\leq$ 2 or 5 kb; (3) *neighboring-variant* set comprising random SNVs located $\leq$ 2 or 5 kb to the positive ones; and (4) *random-gene* set comprising random gene TSSs located $\leq$ 2 or 5 Mb of the eVariants. RegVar models trained previously exhibited an equal or even slightly increased prediction power in these independent negative controls selected from wider genome regions, whereas other methods again showed very limited prediction performance (Figure S15).

We examined the feature importance of the four different models with Gini impurity in the liver datasets (File S1). For *random-variant* and *mirrored-variant* models, the epigenetic patterns of variants were the most important feature sets, while for the *random-gene* model, the epigenetic and sequential profiles of TSSs were the most important feature sets in addition to the distance between variants and TSSs. Notably, for the *neighboring-variant* model, evolutionary and sequential profiles of variants became the most important feature sets (Figure S16). This is expected, as these features could provide information of single-base resolution, which is crucial for distinguishing regulatory variants from adjacent nonfunctional variants.

Characterization of RegVar score distribution across different functional genome regions

We further compared the performance of different RegVar models trained on the GTEx liver eQTLs on each of the negative datasets constructed for the independent exSNP testing set. The results showed that the AUCs of models trained on other types of negative data decreased by 0.031–0.133, 0.041–0.091, and 0.062–0.105 for *random-variant*, *mirrored-variant*, and *neighboring-variant* datasets, respectively, and that the integrated model demonstrated superior capability in all testing datasets in addition to the models trained on the same type of negative data (Figure S17). We then applied the integrated model to all SNVs on chromosome 22 (n = 1,039,985) in 17 types of tissues and measured their regulatory potentials with the corresponding RegVar scores. We calculated the optimal cutoff of RegVar scores by maximizing the sum of specificity and sensitivity. We found that a major proportion (84.5%–94.0%) of the DHS-supported eVariants were correctly classified, and a significant subset (24.6%–39.1%) of background variants were assigned RegVar scores above the cutoffs (**Figure 4**A, Figure S18). This result suggests that a considerable number of variants in the human genome can function as regulatory variants.

To further investigate the distribution of RegVar scores across different functional genome regions, we mapped all annotated variants across 15 chromatin states produced by ChromHMM [22] in the liver. The results showed that variants at active/bivalent promoters and enhancers usually have higher RegVar scores, while variants at repressed and heterochromatin regions usually have lower scores (Figure 4B). This distribution is expected since most variants exert their effects through alteration of key regulatory DNA elements [33]. Additionally, we observed a clear correlation between RegVar scores and SNV-caused loss-of-function ($\Delta Affinity \leq 0$, ANOVA F = 422.6, P = 0) or gain-of-function ($\Delta Affinity \geq 0$, ANOVA F = 23.62, P = 5.52E–11) TFBSs (Figure 4C), which means that the extent of TFBS affinity alteration is positively correlated with the probability of the causing variants being functional.

RegVar shows capability to identify tissue-specific regulatory variants

To evaluate the tissue specificity of the predicted regulatory variants, we applied the integrated RegVar models in all 17

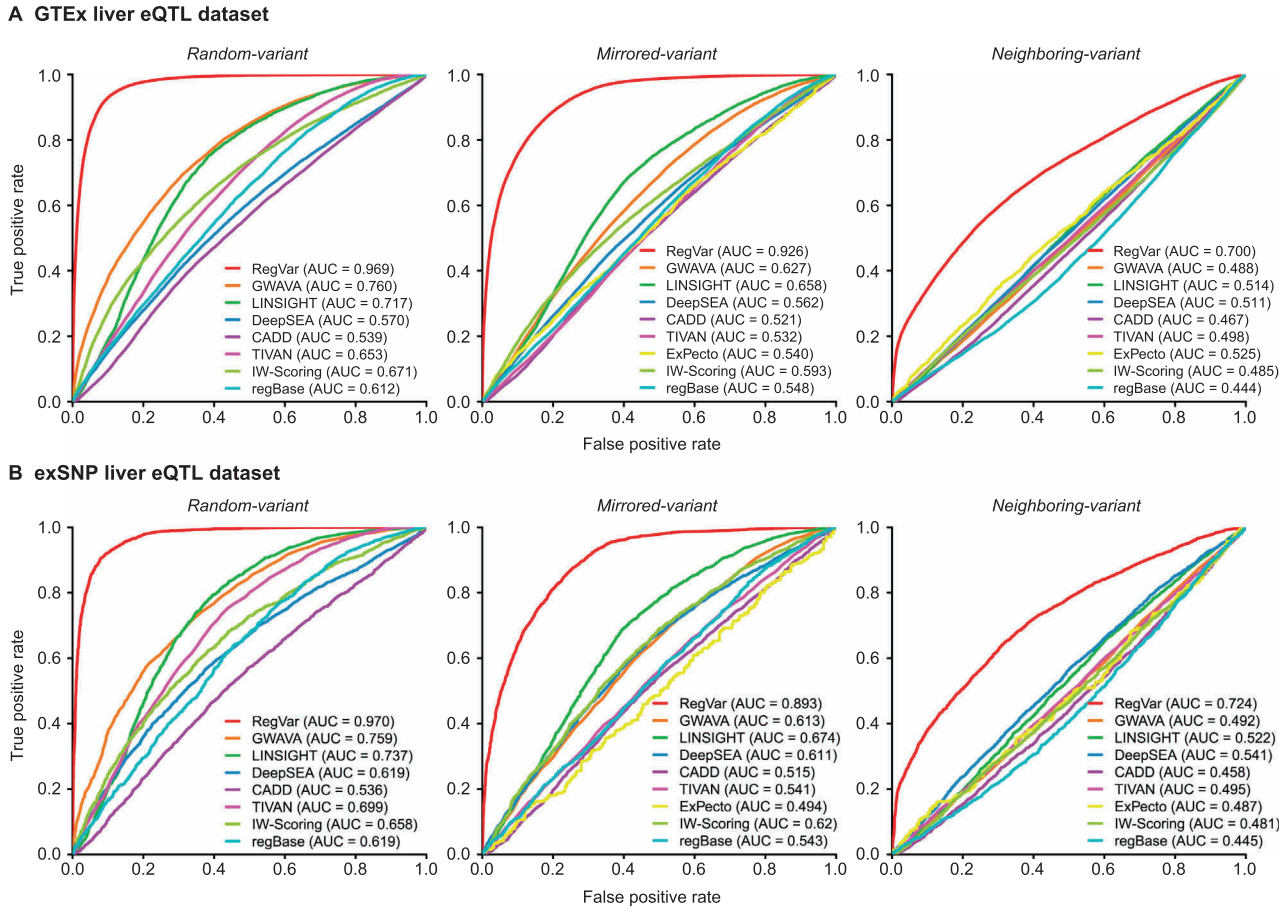


Figure 3 ROC curves of nine computational methods distinguishing regulatory variants from different backgrounds in liver
The results are shown for ROC curves from 10-fold CV experiments in the GTEx liver eQTL dataset (**A**) and from external evaluation experiments in the exSNP liver eQTL dataset (**B**). Negative datasets were from randomly selected variants (*random-variant* sets; left), or matched variants by distance but at the opposite side of the TSS of the eGene (*mirrored-variant* sets; middle), or neighboring variants located adjacent ($\leq$ 1 kb) to the positive ones (*neighboring-variant* sets; right). Negative datasets from randomly selected TSSs (*random-gene* sets) are not shown, since other existing methods did not predict potentially affected genes. Any overlap between the exSNP liver eQTLs and GTEx liver eQTLs and overlap between their corresponding negative sets were removed. ExPecto results are not shown for *random-variant* sets, because it resulted in too few samples for ROC curve analysis. eQTL, expression quantitative trait locus; TSS, transcription start site.

tissue types to randomly selected SNVs (n = 100,000) across the human genome. *K*-means clustering of the RegVar score percentiles of these SNVs identified 22 variant clusters, and one cluster was considered enriched in a specific tissue if it was endowed with a *K*-means center percentile larger than the percentile of the cutoff score in the corresponding tissue. We then identified 8 nonfunctional regulatory variant clusters that were enriched in 0 tissues, 11 tissue-specific regulatory variant clusters that were enriched in 1–3 tissues, and 3 tissue-shared regulatory variant clusters that were enriched in $\geq$ 12 tissues (there was no cluster enriched in 4–11 tissues) (**Figure 5**A, Figure S19). Using four epigenetic marks (DHS, H3K4me1, H3K27ac, and H3K4me3) as hallmarks of chromatin states, we showed that two clusters of tissue-shared variants assigned high RegVar scores (C6, C14) presented active promoter marks (DHS, H3K4me1, H3K27ac, and H3K4me3) across all tissues, indicating that they were enriched at tissue-shared promoters. In contrast, the tissue-shared cluster assigned moderate RegVar scores (C2) presented active enhancer marks (DHS, H3K4me1, and H3K27ac) across all tissues, indicating their enrichment at tissue-shared enhancers. We also found that most of the tissue-specific clusters presented active enhancer marks specifically in the corresponding tissues, indicating their enrichment at tissue-specific enhancers (Figure 5B). These results demonstrate the power of RegVar in measuring the tissue-specific impact of regulatory variants.

RegVar shows competitive performance in prioritization of non-coding pathogenic variants

We further extended the framework of RegVar to prioritize non-coding pathogenic variants. We used a simplified pathogenic RegVar model to learn the features of

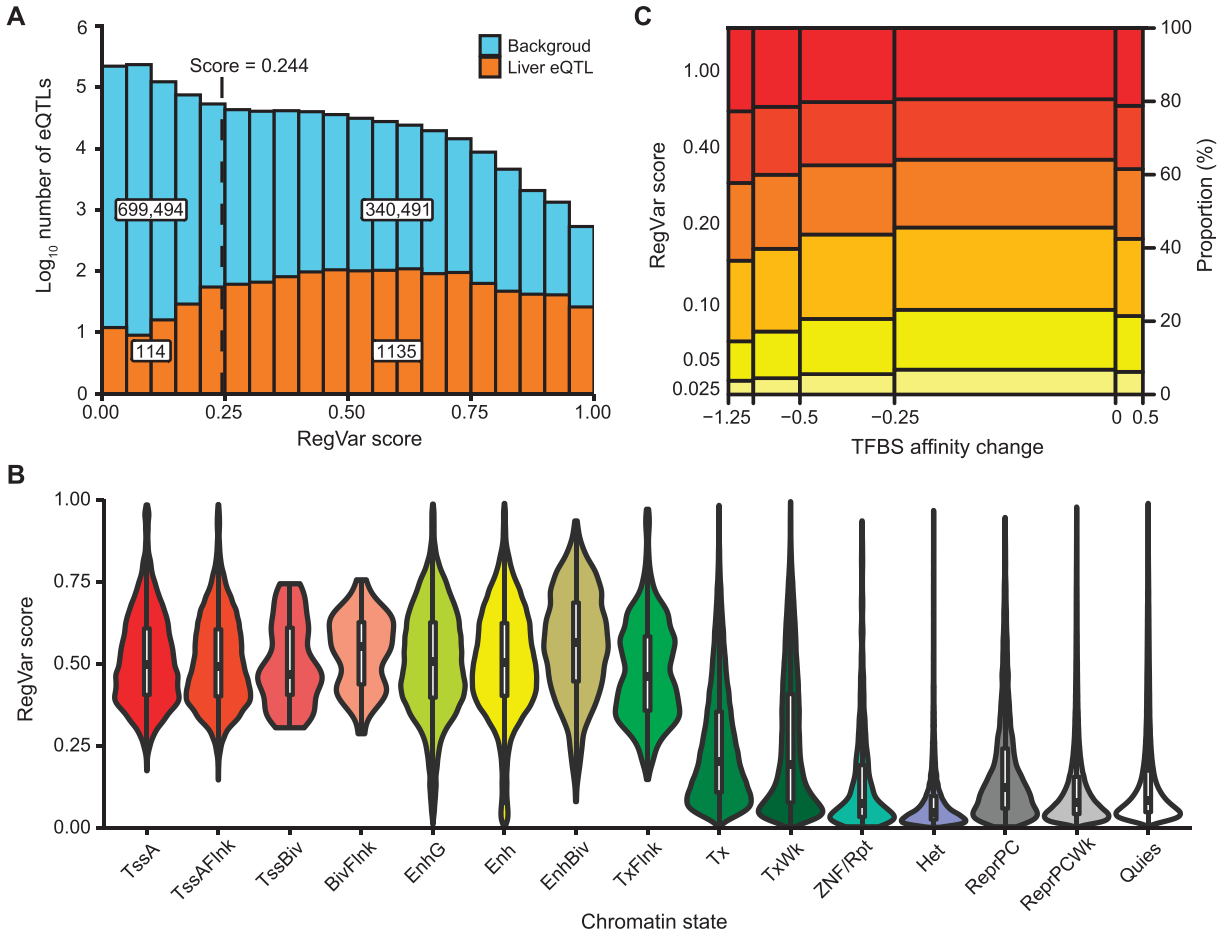


Figure 4 RegVar scores across all variants on chromosome 22 annotated in the integrated liver RegVar model

A. Histogram showing the RegVar score distribution across all SNVs on chromosome 22 (n = 1,039,985) (light blue) and those SNVs in GTEx liver eQTLs (orange). The dashed line indicates the optimal cutoff score in the liver training set. Numbers of variants below or above the cutoff score are embedded. **B.** Violin plots showing the RegVar score distributions across 15 chromatin states. Embedded boxplots indicate medians (center bars) and the first and third quartiles (lower and upper hinges). **C.** Spine plot showing the correlation between RegVar scores and SNV-caused TFBS affinity changes. BivFlnk, flanking bivalent TSS/enhancer; Enh, enhancer; EnhBiv, bivalent enhancer; EnhG, genic enhancer; Het, heterochromatin; ReprPC, repressed PolyComb; ReprPCWk, weak repressed PolyComb; Quies, quiescent/low; TssA, active TSS; TssAFlnk, flanking active TSS; TssBiv, bivalent/poised TSS; Tx, strong transcription; TxFlnk, transcription at gene 5′ and 3′; TxWk, weak transcription; ZNF/Rpt, ZNF gene and repeat; TFBS, transcription factor binding site.

non-coding pathogenic variants collected from the Human Gene Mutation Database (HGMD) [34]. We extracted disease-associated variants from the December 2016 release of the HGMD public dataset. Small indels and variants overlapping any coding sequence (as annotated in RefSeq genes from the UCSC Genome Browser) or essential splice site (as annotated in GWAVA [4]) were filtered out. After mapping the remaining variants to all genomic annotations (File S1), a final set of 2078 disease-associated variants was used as the positive set of non-coding pathogenic variants. For negative datasets, three datasets were constructed: (1) *random-variant* set of random SNVs sampled from the whole genome; (2) *distance-control-variant* set of random SNVs sampled from variants matched to the pathogenic ones by the exact distance-to-nearest TSS (not necessarily near the same TSSs as the pathogenic variants); and (3) *neighboring-variant* set of random SNVs located $\leq$ 1 kb from the pathogenic variants. We selected a sample ratio of 1:10 due to the small sample size of pathogenic variants, and negative variants overlapping any coding sequence or essential splice site were further filtered out. We then constructed the pathogenic RegVar model on those different negative datasets. We found that RegVar demonstrated superior capability in the *random-variant* and *neighboring-variant* sets. The performance of regBase (AUC = 0.879), GWAVA (AUC = 0.874), and IW-Scoring (AUC = 0.871) were comparable to RegVar (AUC = 0.885) in the *random-variant* set, and regBase (AUC = 0.704) was comparable to RegVar (AUC = 0.707) in the *neighboring-variant* set. In the *distance-control-variant* set, regBase exhibited slight outperformance (AUC = 0.845), followed by RegVar (AUC = 0.816) and IW-Scoring (AUC = 0.805) (**Figure 6**). These results demonstrate the competence of the RegVar framework in discriminating between pathogenic and benign variants.

We then explored the feature importance of the aforementioned three models (File S1). We found that sequential profiles were the most important feature set in all three models, illustrating their prominent role in discriminating non-coding

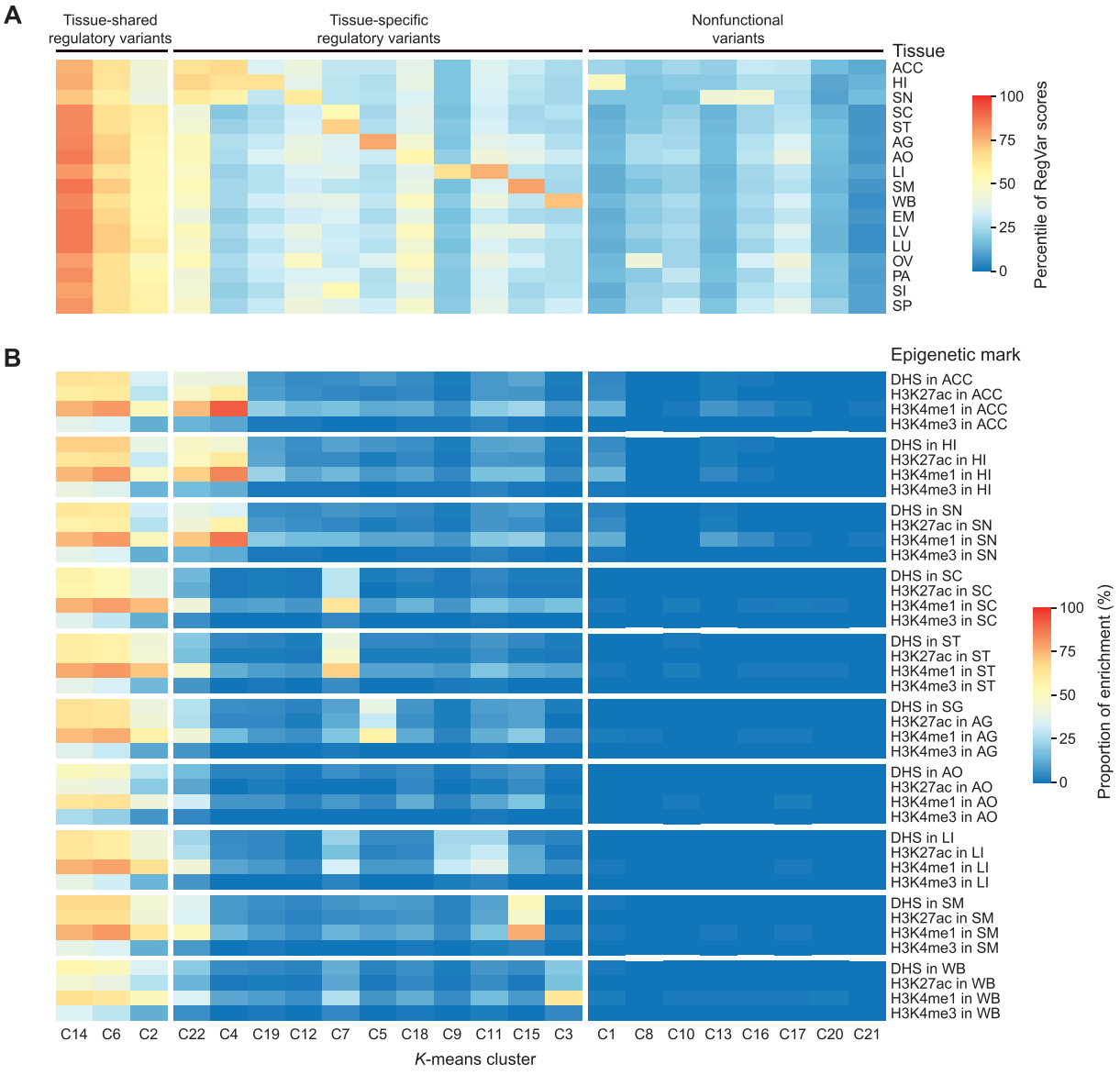


Figure 5 Nonfunctional, tissue-shared, and tissue-specific regulatory variants identified by *K*-means clustering
A. RegVar score percentiles for different clusters of variants (n = 100,000) annotated with the integrated RegVar models in 17 tissues. **B.** Enrichment proportion of different clusters of variants in genome regions with four epigenomic annotations (DHS, H3K4me1, H3K4me3, and H3K27ac) in 10 selected tissues. ACC, anterior cingulate cortex; AG, adrenal gland; AO, aorta; EM, esophagus mucosa; HI, hippocampus; LI, liver; LU, lung; LV, left ventricle; OV, ovary; PA, pancreas; SC, sigmoid colon; SI, small intestine; SM, skeletal muscle; SN, substantia nigra; SP, spleen; ST, stomach; WB, whole blood.

pathogenic variants from different backgrounds. Epigenetic and evolutionary profiles were ranked second in the *random-variant* and *neighboring-variant* models, respectively (Figure S20), which demonstrates their specific facility in separating non-coding pathogenic variants from global and local genome regions, respectively.

Discussion

Non-coding variants play a prominent role in many diseases and complex traits through various intricate mechanisms [35,36]. Nevertheless, variants exert their effects by affecting the expression of specific genes. It is a great challenge to link regulatory variants, especially long distances, to their target genes. Here, we show that by jointly learning the genomic patterns of variants and genes, RegVar provides helpful information for mapping regulatory variants to their target genes. We expect RegVar to contribute to the current limited understanding of the genetic architecture of the human genome and help to uncover novel molecular mechanisms underlying complex traits and diseases.

Numerous methods have been developed for measuring the consequences and importance of non-coding variants. Although differing from each other in the underlying

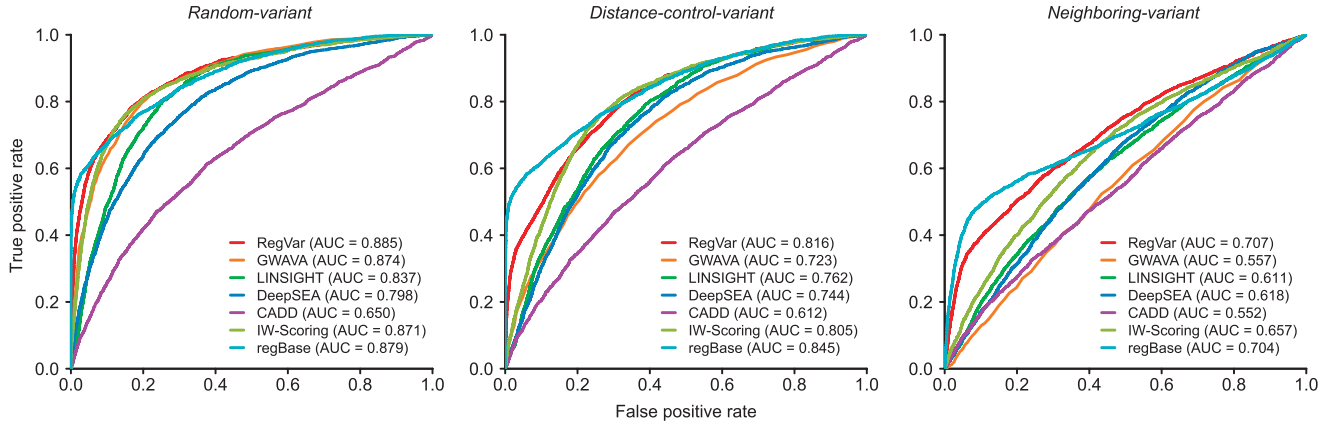


Figure 6 ROC curves of seven computational methods distinguishing non-coding pathogenic variants from different backgrounds
Positive samples were non-coding pathogenic variants (n = 2078) from the HGMD. Negative samples were from randomly selected variants (*random-variant* set; left), or matched variants by the exact distance-to-nearest TSS (not necessarily near the same TSS as each pathogenic variant) (*distance-control-variant* set; middle), or neighboring variants located adjacent ($\leq$ 1 kb) to the positive ones (*neighboring-variant* set; right). Because GWAVA was trained on HGMD non-coding mutations, we filtered out GWAVA training positive variants to evaluate its performance. TIVAN and ExPecto results are not shown because they only provide tissue-specific regulatory variant prioritization scores. HGMD, Human Gene Mutation Database.

assumptions and specific algorithm frameworks, they focused mainly on predicting the pathogenic effect of variants. Therefore, a vast number of non-coding variants with smaller regulatory effects would be neglected. Here, we demonstrated the unique ability of RegVar to prioritize regulatory variants against different backgrounds. We found that in the *random-variant*, *mirrored-variant*, and *random-gene* datasets, RegVar obtained accurate and robust prediction capability; in the *neighboring-variant* dataset, RegVar exhibited relatively weak prediction power but was still superior to existing methods. These results demonstrate RegVar as an integrated model to identify genome-wide regulatory variants, and it may not be suitable for fine-mapping studies in limited regions. By applying RegVar to all SNVs on chromosome 22, we showed that there is a considerable portion of variants across the wide genome showing large probabilities with which to regulate the expression of certain target genes. The reason that they have not been reported may be that their effects are too subtle to be detected, coupled with limited sample sizes and low statistical power.

Code availability

The RegVar online server is freely available at https://regvar.omic.tech/. Downloadable datasets and source code to run RegVar on local personal computers and scripts to generate figures in this study are also provided on the RegVar website.

Competing interests

The authors have declared no competing interests.

CRediT authorship contribution statement

Hao Lu: Methodology, Investigation, Software, Visualization, Writing – original draft, Writing – review & editing. **Luyu Ma:** Methodology, Investigation, Visualization. **Cheng Quan:** Investigation, Software. **Lei Li:** Methodology, Investigation, Visualization. **Yiming Lu:** Conceptualization, Methodology, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Gangqiao Zhou:** Conceptualization, Supervision. **Chenggang Zhang:** Conceptualization, Supervision. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the General Program of the National Natural Science Foundation of China (Grant No. 31771397) and the Beijing Nova Program (Grant No. 20180059).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.08.011.

ORCID

ORCID 0000-0002-8157-4158 (Hao Lu)
ORCID 0000-0003-2907-8410 (Luyu Ma)
ORCID 0000-0003-1859-9683 (Cheng Quan)
ORCID 0000-0002-5100-2124 (Lei Li)
ORCID 0000-0001-8005-2705 (Yiming Lu)
ORCID 0000-0002-4895-5063 (Gangqiao Zhou)
ORCID 0000-0002-4521-3304 (Chenggang Zhang)

References

[1] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337:1190–5.

[2] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.
[3] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5.
[4] Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods 2014;11:294–6.
[5] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931–4.
[6] 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.
[7] Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ, et al. Aberrant gene expression in humans. PLoS Genet 2015;11:e1004942.
[8] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet 2018;19:581–90.
[9] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 2017;101:5–22.
[10] Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. Nature 2016;536:41–7.
[11] Liu L, Sanderford MD, Patel R, Chandrashekar P, Gibson G, Kumar S. Biological relevance of computationally predicted pathogenicity of noncoding variants. Nat Commun 2019;10:330.
[12] Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk. Nat Genet 2018;50:1171–9.
[13] Chen L, Wang Y, Yao B, Mitra A, Wang X, Qin X. TIVAN: tissue-specific *cis*-eQTL single nucleotide variant annotation and prediction. Bioinformatics 2019;35:1573–5.
[14] Li MJ, Li M, Liu Z, Yan B, Pan Z, Huang D, et al. cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. Genome Biol 2017;18:52.
[15] Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell 2016;165:1519–29.
[16] GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204–13.
[17] Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 2003;31:374–8.
[18] Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2016;44:D110–5.
[19] Liefooghe A, Touzet H, Varré JS. Large scale matching for position weight matrices. Lect Notes Comput Sci 2006;4009:401–12.
[20] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2.
[21] Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 2004;5:276–87.
[22] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 2012;9:215–6.
[23] Schmitt A, Hu M, Jung I, Xu Z, Qiu Y, Tan C, et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 2016;17:2042–59.
[24] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21.
[25] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034–50.
[26] Wang J, Ullah AZD, Chelala C. IW-Scoring: an Integrative Weighted Scoring framework for annotating and prioritizing genetic variations in the noncoding genome. Nucleic Acids Res 2018;46:e47.
[27] Zhang S, He Y, Liu H, Zhai H, Huang D, Yi X, et al. regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. Nucleic Acids Res 2019;47:e134.
[28] Yu CH, Pal LR, Moult J. Consensus genome-wide expression quantitative trait loci and their relationship with human complex trait disease. OMICS 2016;20:400–14.
[29] Schulz H, Ruppert AK, Herms S, Wolf C, Mirza-Schreiber N, Stegle O, et al. Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nat Commun 2017;8:1511.
[30] Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci 2014;17:1418–28.
[31] Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of *trans* eQTLs as putative drivers of known disease associations. Nat Genet 2013;45:1238–43.
[32] Brown CD, Mangravite LM, Engelhardt BE, Gibson G. Integrative modeling of eQTLs and *cis*-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genet 2013;9:e1003649.
[33] Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. Principles and methods of *in-silico* prioritization of non-coding regulatory variants. Hum Genet 2018;137:15–30.
[34] Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 2017;136:665–77.
[35] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet 2015;16:197–212.
[36] Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nat Rev Genet 2016;17:93–108.