RegVar: Tissue-specific Prioritization of Noncoding Regulatory Variants

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

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

* Corresponding authors.

E-mail: ylu.phd@gmail.com (Lu Y), zhougq114@126.com (Zhou G), zhangcglab@gmail.com (Zhang C)

Running title: *Lu H et al / Prioritization of Regulatory Variants*

Article title letter count: 65

Running title letter count: 42

Abstract word count: 143

Total word count (from “Introduction” to “Discussion”): 4625

Total figures: 5

Total tables: 0

Total supplementary figures: 20

Total supplementary tables: 7

Total supplementary file: 1

Abstract

Noncoding genomic variants constitute the majority of trait-associated genome variations; however, identification of functional noncoding variants is still a challenge in human genetics, and a method systematically assessing the impact of regulatory variants on gene expression and linking them to potential target genes is still lacking. Here we introduce a deep neural network (DNN)-based computational framework, RegVar, that can accurately predict the tissue-specific impact of noncoding 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 noncoding variants 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 webserver at http://regvar.cbportal.org/.

KEYWORDS: Variant prioritization; Expression regulation; Deep neural network

Introduction

Trait-associated genetic variants usually lie in noncoding genomic regions [1, 2], and interpretation of functional noncoding variants is crucial for revealing the underlying genetic architecture and molecular mechanism of complex traits and diseases. Several methods have been developed to discriminate pathogenic variants from non-pathogenic ones using genomic sequences, functional annotations and evolutionary features, such as CADD [3], GWAVA [4], DeepSEA [5], LINSIGHT [6], etc. A common feature of these methods is that they focus on identifying rare pathogenic variants, which were thought to have stronger impact on human traits and diseases than common variants [7]. However, emerging evidences suggest that the majority 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 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 noncoding 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]. This is because prioritization of regulatory variants is an even greater challenge than that of pathogenic ones. First, regulatory variants generally have weaker impact on gene expression compared to pathogenic ones, so it is more difficult to discriminate them from background, especially from adjacent non-functional variants sharing similar epigenetic marks. Second, it is challenging to link regulatory variants to their target genes, which can be located far away from its 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], which, however, have their 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 a variant's regulatory probability, but it was trained with promoter-proximal variants, which may introduce potential biases when applied to genome-wide variants prioritization [13]. Considering the vast majority of regulatory variants 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 remains an urgent need.

Here we introduce a deep neural network (DNN)-based approach, RegVar, for the genome-wide assessment of the regulatory impact of noncoding variants on gene expression. RegVar has several key features: (i) it can predict both common and rare regulatory variants by learning their genomic characteristics from massive variant-gene associations in an unbiased manner; (ii) it predicts not only regulatory variants but their target genes by jointly learning the genomic patterns of both variants and genes and the chromatin interactions between them; (iii) it predicts the tissue-specific effects of variants by training models in multiple tissues with respective genomic patterns; and (iv) it can achieve excellent prediction accuracy by utilizing large training sets and deep learning algorithm. We show that RegVar outperforms existing prioritization methods in identifying regulatory variants and noncoding pathogenic variants from different backgrounds in various tissues. RegVar is available as a webserver at http://regvar.cbportal.org/.

Materials and methods

Datasets

To construct the positive datasets, significant eVariant-eGene associations in 17 human tissues, which were also incorporated in the Roadmap Epigenomics projects [2], were obtained from GTEx V7 release [16] (**Figure 1** and Table S1). Single-nucleotide variants (SNVs)-gene associations were selected and further filtered by removing eVariants not marked by DNase I hypersensitive sites (DHSs) annotations, which was demonstrated to be a key epigenetic marker of causal variants [16]. Associations in sex chromosomes were also removed. For tissues of which the numbers of significant associations exceed 100,000 (esophagus mucosa, lung, skeletal muscle, and whole blood), we randomly selected 100,000 associations, as we found a larger size cannot improve model performance (Figure S1). The final number of positive associations for each tissue was shown in Table S1. For negative datasets, four datasets were constructed, including: (i) *random-variant* set of shuffled SNV-gene pairs where eVariants were replaced by random SNVs located <=1 Mb from the eGene TSS; (ii) *mirrored-variant* set of shuffled pairs where eVariants were replaced by random SNVs located at similar distance (error <= 1kb) but the opposite side of the eGene TSS; (iii) *neighboring-variant* set of shuffled pairs where eVariants were replaced by random SNVs located adjacent (<= 1 kb) to the positive ones; (iv) *random-gene* set of shuffled SNV-gene pairs where eGenes were replaced by gene TSSs located <= 1 Mb of the eVariants. We selected a maximum distance at 1 Mb between SNV and TSS in the datasets (i) and (iv), for it was observed that all positive SNV-TSS pairs had a distance less than 1 Mb (Figure S2). To determine the ratio between positive and negative datasets, we assessed different ratios, including 1:1, 1:2, 1:3, 1:5, and 1:10, and found there was no significant difference of performances among five models (Figure S3). Thus, we selected a ratio of 1:1 between the positive and negative datasets to efficiently train the models. Variants in negative datasets were selected from the dbSNP build 146 data after removing the shared variants between GTEx and dbSNP datasets. Since eQTL variants are biased toward high frequency variants (Figure S4), 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 GTEx liver dataset, by the same strategy as the first three control datasets (i-iii) 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 TFs from the TRANSFAC [17] (523 TFs) and JASPAR [18] (79 TFs) databases. TFMscan [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*-value < 4.5E-5 was used to determine significant motifs. Variants located within these motifs were determined using BEDTools [20]. The TFBS affinity were calculated as described [21]. Specifically, the corrected probabilities of observing a given nucleotide in a specific locus were calculated as follows:

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

Where b represents one specific base among A, T, C, and G, i is the index of the site, $f_{b,i}$ is the counts of base b in site i, N is the sum of 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

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

Hence, the corresponding PWM can be constructed as:

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

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

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

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

$$\Delta Affinity = \frac{Affinity_A - Affinity_R}{w} \quad (5)$$

where $Affinity_R$ and $Affinity_A$ are evaluated binding affinity 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 interacted regions (FIREs) annotations from Hi-C study [23].

Evolutionary profiles

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

vertebrate phyloP and phastCons scores based on the 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 if SNV or TSS 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 scaled by

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

and $\Delta Affinity$ scores 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 non-linear 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 used rectified linear unit function (ReLU) as the activation function. Exactly, one fully connected layer computes

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

where X is the input, W is the weight matrix, b is the bias, and ReLU represents 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 non-regulatory variant on the specific gene, with scaled probability ranging from 0 to 1 using

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

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

$$objective = -\frac{1}{N}\sum_{i=1}^{N}\left[Y_i * \log(f(X_i)) + (1-Y_i) * \log(1-f(X_i))\right] \tag{12}$$

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 optimal search of hyper-parameters 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 AUC in each of the four models (Table S3-S6).

All training programs were written in Python language, using a deep neural network implementation from the TensorFlow library.

Model comparison

We used the average receiver operating characteristic (ROC) curves computed from 10-fold cross-validation to evaluate model performances. Specifically, each dataset comprising of 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 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 *random-gene* dataset was excluded as the existing methods didn't give prediction on potentially affected genes. In addition, ExPecto was excluded from the *random-variant* and *random-gene* models evaluation, because it focused on promoter-proximal variants thus resulted 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] eQTLs studies, and blood eQTLs from Westra [31] eQTLs meta-analysis to evaluate performances of trained models. We identified all SNV-TSS pairs and removed those overlapping with liver, hippocampus and whole blood eQTLs in 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. Also, 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.

Besides, we evaluated prediction capabilities of different methods on the liver and blood eQTLs data from Brown eQTLs analysis [32], which have been used to test the performance of TIVAN and regBase. We downloaded the compiled positive and negative sets for Brown eQTLs data from the regBase website and compared

performance of different methods on datasets in liver and blood (all testing datasets are summarized in Table S7 and see Supplementary Methods for more details about data processing).

RegVar score distribution

For each tissue, we trained an integrated RegVar model by pooling four negative datasets to take all conditions together. 17 integrated RegVar models were applied to annotate all possible SNV-gene pairs in chromosome 22. For each SNV in 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 mapped with all kinds of features, 65,844,726 sample pairs of 1,039,985 different SNVs were left and 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 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 mapped with corresponding features, 3,703,900 sample pairs were left. 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 *kmeans* function in R language, was applied on the matrix to get 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

Prioritization of regulatory variants in 17 human tissues

To explore the influence of DHS filter on RegVar's prediction capability , we first compared performance of models built from positive datasets without any filter, with the DHS filter, with the ATAC-seq filter, and both positive and negative datasets with the DHS filter in the liver dataset (Supplementary Methods), and found that models built from the positive dataset with the DHS filter showed the most robust performance in discriminating regulatory variants form different backgrounds (Figure S5). We then utilized the 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** and Table S1). This result demonstrated RegVar could reliably discriminate positive regulatory variants from different negative backgrounds. We then evaluated the performances 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* and 0.573-0.677 for *mirrored-variant* datasets, yet still much lower than that of RegVar (0.957-0.969 for *random-variant* and 0.884-0.945 for

315 *mirrored-variant* sets), while other five methods failed to show significant power in
316 distinguishing regulatory variants (**Figure 3**, Figure S6 and S7). For the
317 *neighboring-variant* set, which is more challenging, none of the existing methods
318 made valid predictions, compared to an AUC of 0.694-0.700 for RegVar. We
319 additionally evaluate the prediction results of different methods by precision-recall
320 curves (PRC), and still RegVar showed the superior performance against other
321 methods (Figure S8, S9, and S10). After controlled for MAF, RegVar showed
322 comparable prediction capabilities in discriminating eQTLs from MAF-matched
323 benign variants, as demonstrated in liver samples (Figure S11), although with a slight
324 decrease in the *neighboring-variant* set, and the other eight methods still showed low
325 prediction capabilities as before.

326 To further confirm the results, we curated another three publicly available eQTL
327 datasets of liver, hippocampus, and whole blood, from the exSNP website [28],
328 Schulz [29] and Ramasamy [30] eQTL studies, and Westra [31] eQTL meta-analysis,
329 respectively, as independent testing sets. We found RegVar models trained with
330 GTEx datasets achieved almost equally accurate predictions in the three independent
331 testing sets, while all other methods still didn't show any obvious predictive powers in
332 the independent datasets (Figure 3, Figure S6-S10). To assess the robustness of
333 RegVar on imbalanced datasets, we then constructed independent validation sets for
334 liver eQTLs form the exSNP database with the sample ratio at 1:1, 1:2, 1:3, 1:5, and
335 1:10. We found that RegVar trained on GTEx datatsets at the 1:1 ratio showed robust
336 performance on both balanced and imbalanced datasets (Figure S12). In addition, we
337 also evaluated performance of different methods on the Brown eQTL data in liver and
338 blood, which have been used as testing data for regBase and TIVAN. Results showed
339 that in both tissues, RegVar trained on GTEx datasets showed comparable
340 performance (AUC = 0.858 and 0.901 for liver and blood sets, respectively) with
341 regBase (AUC = 0.883 and 0.89 for liver and blood sets, respectively, equal to the
342 AUCs reported in the regBase paper. We also showed that RegVar models trained on

the Brown eQTL data achieved even higher AUCs on both datasets (AUC = 0.952 and 0.945 for liver and blood sets, respectively) compared with other methods (Figure S13 and S14).

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: (i) *random-variant* set comprising of random SNVs located <= 2 and 5 Mb from the eGene TSS; (ii) *mirrored-variant* set comprising of random SNVs with a distance error <= 2 and 5 kb; (iii) *neighboring-variant* set comprising of random SNVs located <= 2 and 5 kb to the positive ones; (iv) *random-gene* set comprising of random gene TSSs located <= 2 and 5 Mb of the eVariants. RegVar models trained before exhibited equal, or even slightly increased, prediction power in these independent negative controls selected from wider genome regions, whereas other methods again showed vary limited prediction performances (Figure S15). Altogether, these results demonstrated the outstanding performance of RegVar on predicting regulatory impact of noncoding variants.

We examined the feature importance of the four different models with Gini impurity in liver (Supplementary Methods). 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 TSS were the most important feature sets besides the distance between variant and TSS. 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 non-functional ones.

RegVar score distribution

We further compared performances of different models trained on the GTEx liver eQTLs on each of negative datasets constructed for the independent exSNP testing set. Results showed that AUCs of models from other types of negative data decreased by 0.031-0.133, 0.041-0.091, 0.062-0.105 for *random-variant*, *mirrored-variant*, and *neighboring-variant* datasets, respectively, and that the integrated model obtained superior capability in all testing datasets besides the models trained on the same type of negative data (Figure S17). We then applied the integrated model to all SNVs in 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 with RegVar scores above the cutoffs (**Figure 4A** and Figure S18). This result suggests that a considerable portion of variants in 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 liver and 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 4C**). This is expected since most variants exert their effects through alteration of key regulatory DNA elements [33]. Also, we observed a clear correlation between RegVar scores and SNV-caused loss-of-function (*ΔAffinity* <= 0, ANOVA F = 422.6, P = 0) or gain-of-function (*ΔAffinity* >= 0, ANOVA F = 23.62, P = 5.52E-11) of TFBSs (**Figure 4B**), which means the extents of TFBS affinity alteration are positively correlated with the probabilities of the causing variants to be functional.

Tissue specificity of RegVar scores

To evaluate the tissue specificity of the predicted regulatory variants, we applied the integrated models in all 17 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 to be 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, 11, and 3 clusters of non-functional, tissue-specific regulatory, and tissue-shared regulatory variants, from clusters which were enriched in 0, 1-3, and >=12 tissues (there was no cluster enriched in 4-11 tissues), respectively (**Figure 5A** and 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 with high RegVar scores (C6, C14) presented active promoter marks (DHS, H3K4me1, H3K27ac, and H3K4me3) across all tissues, indicating they were enriched at tissue-shared promoters. In contrast, the tissue-shared cluster assigned with 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.

Prioritization of noncoding pathogenic variants in HGMD

We further extended the framework of RegVar to prioritize noncoding pathogenic variants. We used a simplified pathogenic RegVar model to learn the features of noncoding pathogenic variants collected from the Human Gene Mutation database (HGMD) [34]. We extracted disease-associated variants from the December 2016 release of HGMD public dataset. Small indels and variants overlapping any coding sequence (as annotated in RefSeq genes from the UCSC Genome Browser) or

426 essential splice sites (as annotated in GWAVA [4]) were filtered out. After mapping
427 remaining variants to all genomic annotations (Supplementary Methods), a final set of
428 2078 disease-associated variants were used as the positive set of pathogenic
429 noncoding variants. For negative datasets, three datasets were constructed, including
430 (i) *random-variant* set of random SNVs sampled from the whole genome; (ii)
431 *distance-control-variant* set of random SNVs sampled from variants matched to the
432 pathogenic ones by the exact distance-to-nearest TSS (not necessarily near the same
433 TSSs as the pathogenic variants); (iii) *neighboring-variant* set of random SNVs
434 located <= 1 kb from the pathogenic ones. We selected a sample ratio at 1:10 due to
435 the small sample size of pathogenic variants, and negative variants overlapping any
436 coding sequence or essential splice sites were further filtered out. We then constructed
437 the pathogenic RegVar model on those different negative datasets. We found that
438 RegVar obtained superior capability in the *random-variant* and *neighboring-variant*
439 sets. The performances of regBase (AUC = 0.879), GWAVA (AUC = 0.874),
440 IW-scoring (AUC = 0.871) were comparable to RegVar's (AUC = 0.885) in the
441 *random-variant* set, and regBase (AUC = 0.704) was comparable to RegVar (AUC =
442 0.707) in the *neighboring-variant* set. In the *distance-control-variant* set, regBase
443 exhibited slight outperformance (AUC = 0.845), followed by RegVar (AUC = 0.816)
444 and IW-scoring (AUC = 0.805) (**Figure 6**). These results demonstrated the
445 competence of the RegVar framework in discriminating between pathogenic and
446 benign variants.

447 We then explored the feature importance of the above three models
448 (Supplementary Methods). We found that sequential profiles were the most important
449 feature set in all three models, illustrating their prominent role in discriminating
450 noncoding pathogenic variants from different backgrounds; epigenetic and
451 evolutionary profiles were ranked second in the *random-variant* and
452 *neighboring-variant* models, respectively (Figure S20), which demonstrated their

specific facility in separating noncoding pathogenic variants from a global and local genome regions, respectively.

Discussion

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

A number of methods have been developed for measuring the consequence and importance of noncoding variants. Though differing from each other in the underlying intuitions and specific algorithm frameworks, they mainly focused on predicting the pathogenic effect of variants. Therefore a vast number of noncoding 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 still superior to existing methods. These results demonstrate RegVar as an integrated model to identify genome-wide regulatory variants, and it may be not suitable for fine-mapping studies in limited regions. Applying RegVar to all SNVs in chromosome 22, we show 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

480 they have not been reported may be that their effects are too subtle to be detected,
481 coupled with limited sample sizes and low statistical power.
482

Code availability

The RegVar online server is freely available at http://regvar.cbportal.org/. Downloadable datasets and source code to run RegVar on local personal computers and scripts to generate figures in the manuscript are also provided at the RegVar website.

CRediT author 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.

Competing interests

The authors have declared no competing interests.

Acknowledgements

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).

Authors' ORCID IDs

0000-0002-8157-4158 (Hao Lu)

0000-0003-2907-8410 (Luyu Ma)

0000-0003-1859-9683 (Cheng Quan)

511 0000-0002-5100-2124 (Lei Li)

512 0000-0001-8005-2705 (Yiming Lu)

513 0000-0002-4895-5063 (Gangqiao Zhou)

514 0000-0002-4521-3304 (Chenggang Zhang)

515

516

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 GR, 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 Y-F, 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. 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] Consortium G. 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 C-y, 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 2015;44:D110-D5.

[19] Liefooghe A, Touzet H, Varré J-S. Large scale matching for position weight matrices. Annual Symposium on Combinatorial Pattern Matching 2006:401-12.

[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 AD, Hu M, Jung I, Xu Z, Qiu Y, Tan CL, 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, Dayem Ullah AZ, 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 C-H, 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 A-K, 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 H-J, 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. 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.

Figure legends

Figure 1 A flowchart showing the workflow of RegVar

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

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

Figure 3 ROC curves of nine computational methods distinguishing regulatory variants from different backgrounds in liver

Results are shown for ROC curves from 10-fold cross-validation experiments in the GTEx liver eQTL dataset (top) and from external evaluation experiments in the exSNP liver eQTL dataset (bottom). Negative datasets were from either random selected variants (*random-variant* sets) (left), matched variants by distance but at the opposite side of the eGene TSS (*mirrored-variant* sets) (middle), or neighboring variants located adjacent (<= 1 kb) to the positive ones (*neighboring-variant* sets) (right). Negative datasets from random selected TSSs (*random-gene* sets) are not shown since other existing methods didn't give prediction on potentially affected genes. Any overlap between the exSNP liver eQTLs and GTEx 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.

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

A. Histogram showing the RegVar scores distribution across all SNVs in chromosome 22 (N = 1,039,985) (lightblue) and SNVs in GTEx liver eQTLs (orange). Dashed line indicates the optimal cutoff score in liver training set. Numbers of variants blow or

above the cutoff score are embedded. **B**. Spine plot showing the correlation between RegVar scores and SNV-caused TFBS affinity changes. **C**. Violin plots showing the RegVar score distributions across 15 chromatin states (BivFlnk, flanking bivalent TSS/enhancers; Enh, enhancers; EnhBiv, bivalent enhancers; EnhG, genic enhancers; 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/Rpts, ZNF genes & repeats). Embedded boxplots indicate medians (center bars), and the first and third quartiles (lower and upper hinges).

Figure 5 Tissue-shared and tissue-specific regulatory variants and non-functional variants identified in K-means clustering

A. RegVar score percentiles for different clusters of variants (N = 100,000) annotated with the integrated RegVar models in 17 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). **B**. Enrichment proportion of different clusters of variants in genome regions with four epigenomic annotations (DHS, H3K4me1, H3K4me3, and H3K27ac) in 10 selected tissues.

Figure 6 ROC curves of seven computational methods distinguishing noncoding pathogenic variants from different backgrounds

Positive samples were from the of HGMD noncoding pathogenic variants (N = 2078). Negative samples were from either random selected variants (*random-variant* set) (left), 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

668 neighboring variants located adjacent (<= 1 kb) to the positive ones
669 (*neighboring-variant* set) (right). Because GWAVA was trained on the HGMD
670 noncoding mutations, we filtered out GWAVA training positive-variants in evaluating
671 its performance. TIVAN and ExPecto results are not shown because they only
672 provides tissue-specific regulatory variants prioritization scores.
673

674 ## Supplementary material

675 **Figure S1 AUCs of RegVar models with different sample sizes**

676 Positive samples were from the GTEx skeletal muscle eQTLs. Sample sizes were set
677 from 20,000 to 180,000, step by 20,000, and the full sample size (N = 204,124).
678

679 **Figure S2 Histogram showing the distance between variant loci and TSS**

680 Results are shown for samples from the GTEx eQTLs from 17 tissues (**A-Q**).
681

682 **Figure S3 ROC curves for 10-fold cross-validation experiments at different**
683 **ample ratios between positive and negative datasets**

684 Positive samples were from the GTEx liver eQTLs. Negative samples were selected at
685 the ratio of 1:1, 1:2, 1:3, 1:5, and 1:10 between positive (P) and negative (N) datasets.
686

687 **Figure S4 MAF distribution (A) and common and rare variant proportion (B) in**
688 **dbSNP variants and GTEx eQTLs**
689

690 **Figure S5 ROC curves of RegVar models trained on samples with different**
691 **filters**

692 Results are shown for models trained on the original GTEx liver eQTLs (N = 268,673)
693 with all feature sets as predictors and with DHS peaks alone as a predictor, RegVar
694 models trained on the GTEx liver eQTLs filtered by DHS peaks (N = 41,636) and

filtered by ATAC profiles (N = 3731), and RegVar models trained on both the GTEx liver eQTLs and controls filtered by DHS peaks.

Figure S6 ROC curves of nine computational methods distinguishing regulatory variants from different backgrounds in hippocampus

Results are shown for ROC curves from 10-fold cross-validation experiments in the GTEx hippocampus eQTL dataset (top) and from external evaluation experiments in Schulz and Ramasamy hippocampus eQTL dataset (bottom). Plots are similar to Figure 3.

Figure S7 ROC curves of nine computational methods distinguishing regulatory variants from different backgrounds in whole blood

Results are shown for ROC curves from 10-fold cross-validation experiments in the GTEx whole blood eQTL dataset (top) and from external evaluation experiments in Westra whole blood eQTL dataset (bottom). Plots are similar to Figure 3.

Figure S8 PRC curves of nine computational methods distinguishing regulatory variants from different backgrounds in liver

Results are shown for PRC curves for the same result as Figure 3.

Figure S9 PRC curves of nine computational methods distinguishing regulatory variants from different backgrounds in hippocampus

Results are shown for PRC curves for the same result as Supplementary Figure S6

Figure S10 PRC curves of nine computational methods distinguishing regulatory variants from different backgrounds in whole blood

Results are shown for PRC curves for the same result as Supplementary Figure S7

Figure S11 ROC curves of nine computational methods distinguishing regulatory variants in liver from MAF-matched variants

Positive samples were from the GTEx liver eQTLs (*N* = 41,636). Negative samples in three control datasets were randomly selected from MAF-matched variants to the positive ones. Plots are similar to Figure 3.

Figure S12 ROC curves of RegVar distinguishing regulatory variants from different backgrounds in liver at different sample ratio

Positive samples were from the exSNP liver eQTL dataset. Negative samples were selected at the ratio of 1:1, 1:2, 1:3, 1:5, and 1:10 between positive (P) and negative (N) datasets.

Figure S13 ROC curves of nine computational methods distinguishing regulatory variants in Brown liver eQTLs

Positive and negative samples were from the Brown liver eQTLs complied by regBase.

Figure S14 ROC curves of nine computational methods distinguishing regulatory variants in Brown blood eQTLs

Positive and negative samples were from the Brown blood eQTLs complied by regBase.

Figure S15 ROC curves of nine computational methods distinguishing regulatory variants in the exSNP liver eQTLs from different backgrounds

Positive samples were from the exSNP liver eQTL dataset (*N* = 4307). Negative samples in three control datasets were randomly selected from wider genome regions: (i) *random-variant* set comprising of random SNVs located <= 2 (Top) and 5 (bottom) Mb from the eGene TSS; (ii) *mirrored-variant* set comprising of random SNVs with a

751 distance error <= 2 (Top) and 5 (bottom) kb; (iii) *neighboring-variant* set comprising
752 of random SNVs located <= 2 (Top) and 5 (bottom) kb to the positive ones. Result for
753 *random-gene* set comprising of random gene TSSs located <= 2 and 5 Mb of the
754 eVariants was not shown since other existing methods didn't give prediction on
755 potentially affected genes.
756
757 **Figure S16 Barplots showing the relative Gini importance for different models**
758 Results are shown for models trained on *random-variant* (**A**), *mirrored-variant* (**B**),
759 *neighboring-variant* (**C**), and *random-gene* (**D**) datasets of the GTEx liver eQTLs (*N*
760 = 41,636). All features were divided into seven groups: distance between variant and
761 TSS (Distance), sequential profiles of variant (VarSeq) and TSS (TSSSeq), epigenetic
762 profiles of variant (VarEpi) and TSS (TSSEpi), evolutionary profiles of variant
763 (VarEvo) and TSS (TSSEvo). Error bars represents the standard error averaged over
764 10 times.
765
766 **Figure S17 ROC curves of different RegVar models distinguishing regulatory**
767 **variants in the exSNP liver eQTLs from each of the negative datasets**
768 Positive samples were from the exSNP liver eQTLs (*N* = 4307). Models based on the
769 *random-gene* dataset was not evaluated in this analysis, because it aimed to identify
770 potential eGenes for each eVariant and wasn't suitable for prioritization of positive
771 variants.
772
773 **Figure S18 Histogram showing the RegVar scores distribution**
774 Results are shown for all SNVs in chromosome 22 (lightblue) and SNVs in the GTEx
775 eQTLs (orange). Variants were annotated in integrated RegVar models in 16 tissues
776 (**A-P**). Dashed line indicates the optimal cutoff scores in the corresponding training
777 set. Numbers of variants blow or above the cutoff scores are embedded.
778

Figure S19 Identification of tissue-shared/tissue-specific cluster of variants

An orange cell indicate the cluster was endowed with an K-means center percentile larger than the percentile of the RegVar cutoff score in the corresponding tissue, and an skyblue cell if not. 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.

Figure S20 Barplots showing the relative Gini importance for different pathogenic RegVar models

Results are shown for models trained on *random-variant* (**A**), *distance-control-variant* (**B**), and *neighboring-variant* (**C**) datasets of HGMD noncoding pathogenic variants (N = 2,078). All features were divided into four groups: distance to the nearest TSS (Distance), and sequential (Seq), epigenetic (Epi), and evolutionary (Evo) profiles of the variant. Error bars represents the standard error averaged over 10 times.

Table S1 Sample sizes and AUCs of the 4 training sets in 17 tissues

Table S2 Summary of genomic features used by RegVar

Table S3 AUCs under different learning rates and dropout proportions in the liver *random-variant* set

Table S4 AUCs under different learning rates and dropout proportions in the liver *mirrored-variant* set

Table S5 AUCs under different learning rates and dropout proportions in the liver *neighboring-variant* set

807

808 **Table S6 AUCs under different learning rates and dropout proportions in the**

809 **liver *random-gene* set**

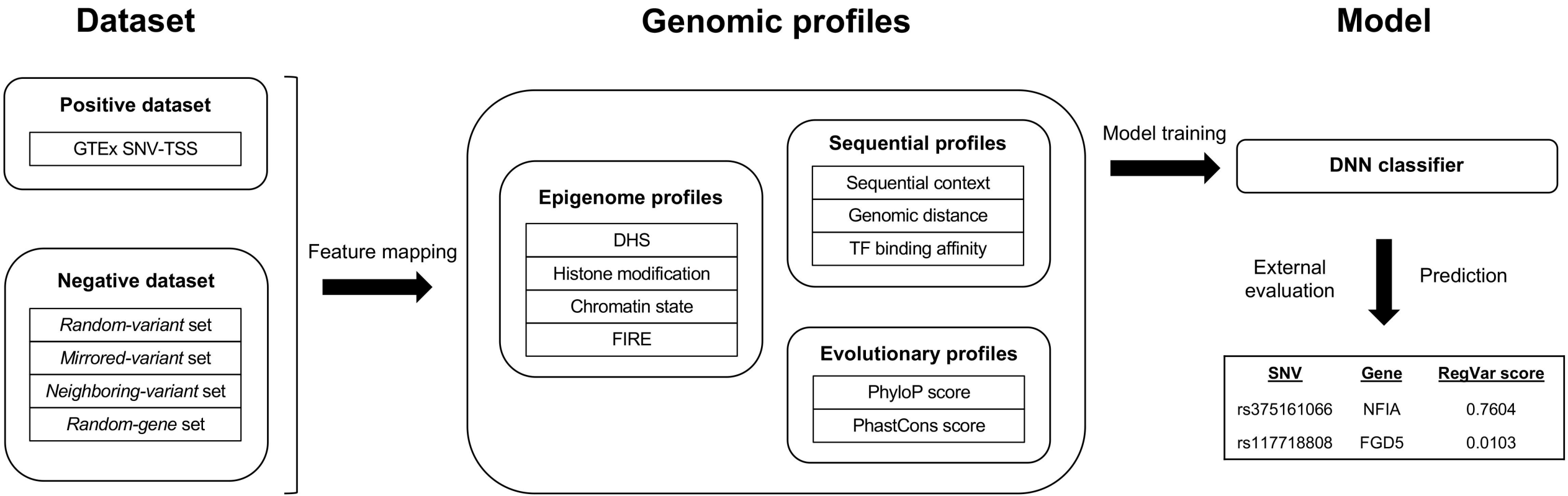
810

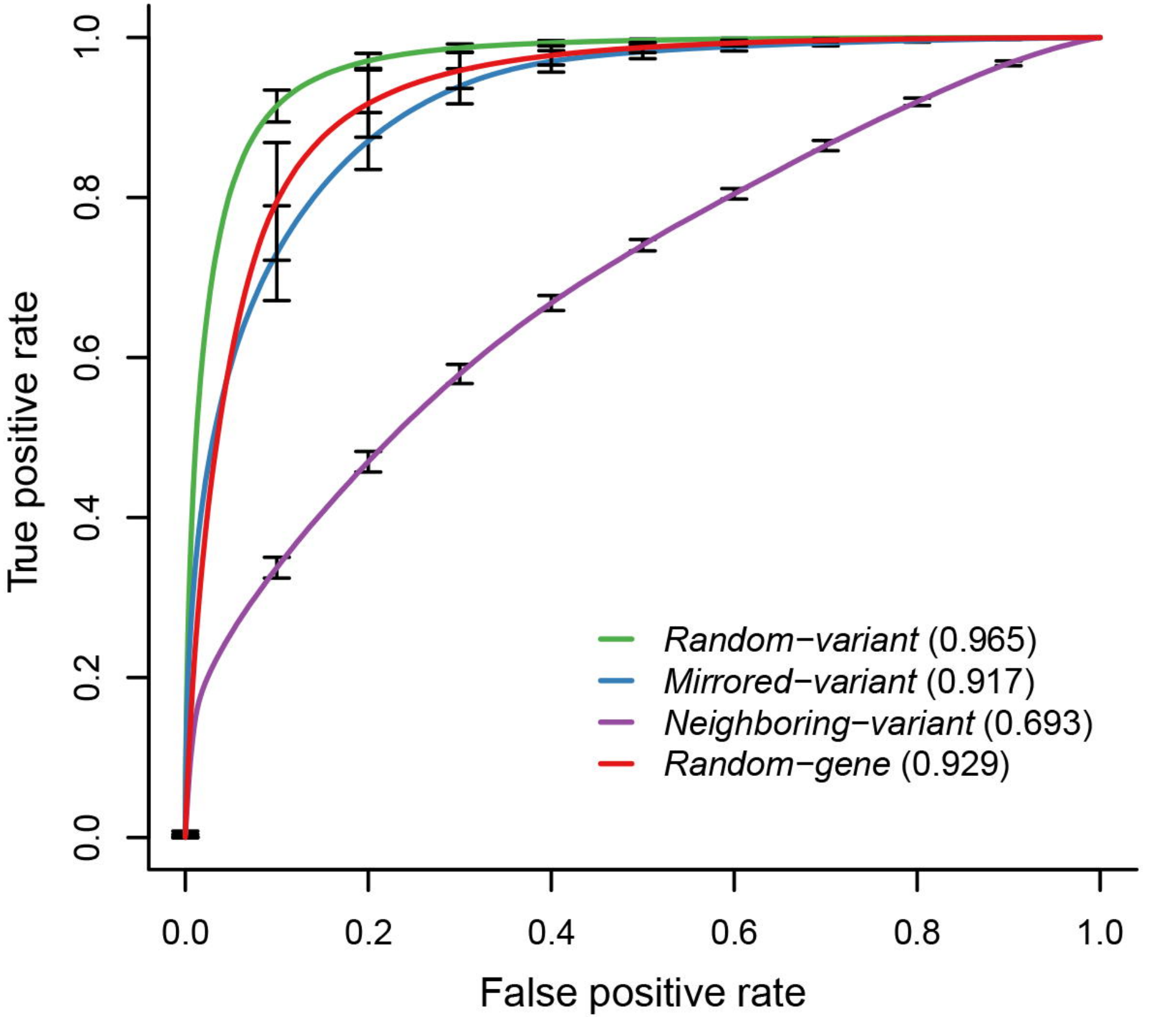
811 **Table S7 Testing sets used for evaluation of RegVar and other methods**

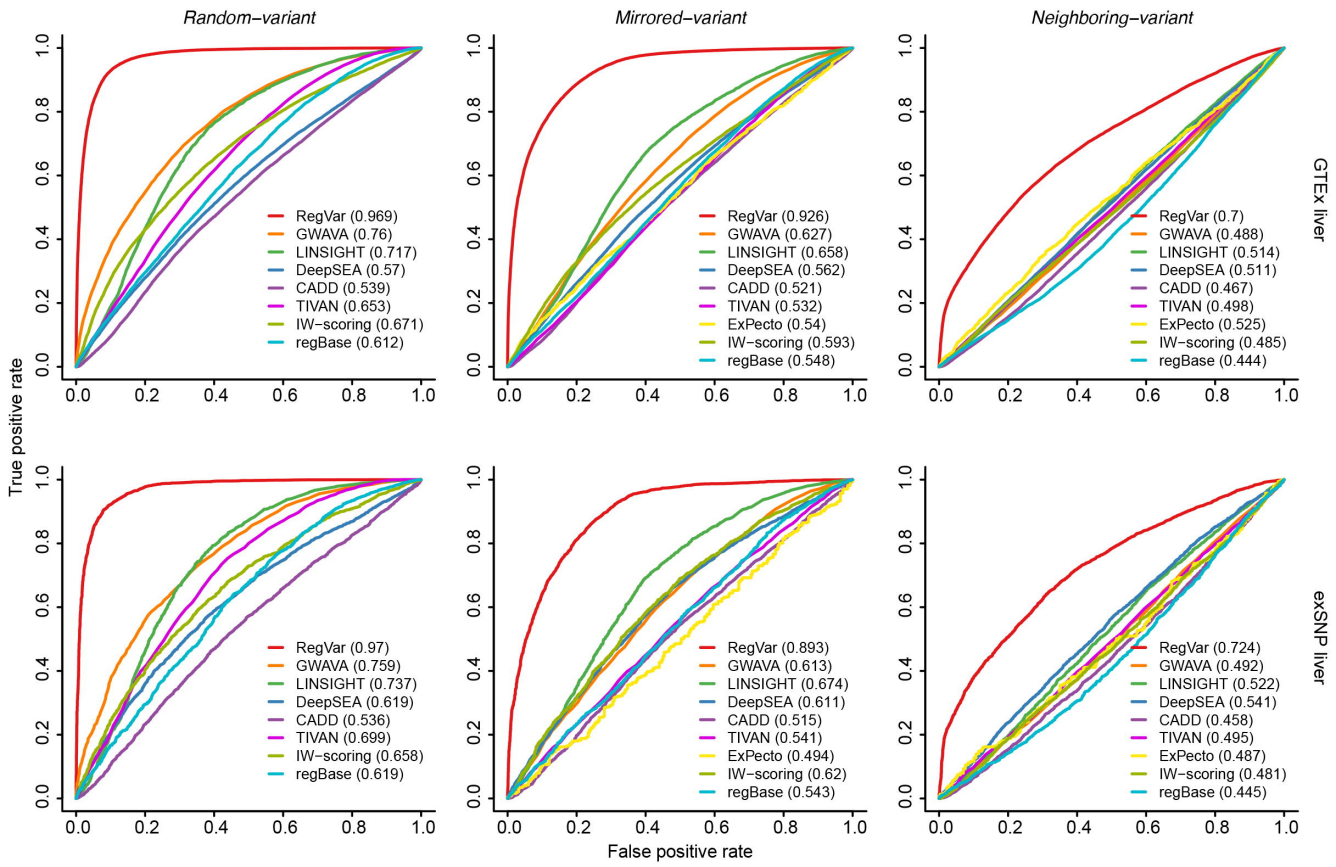
812

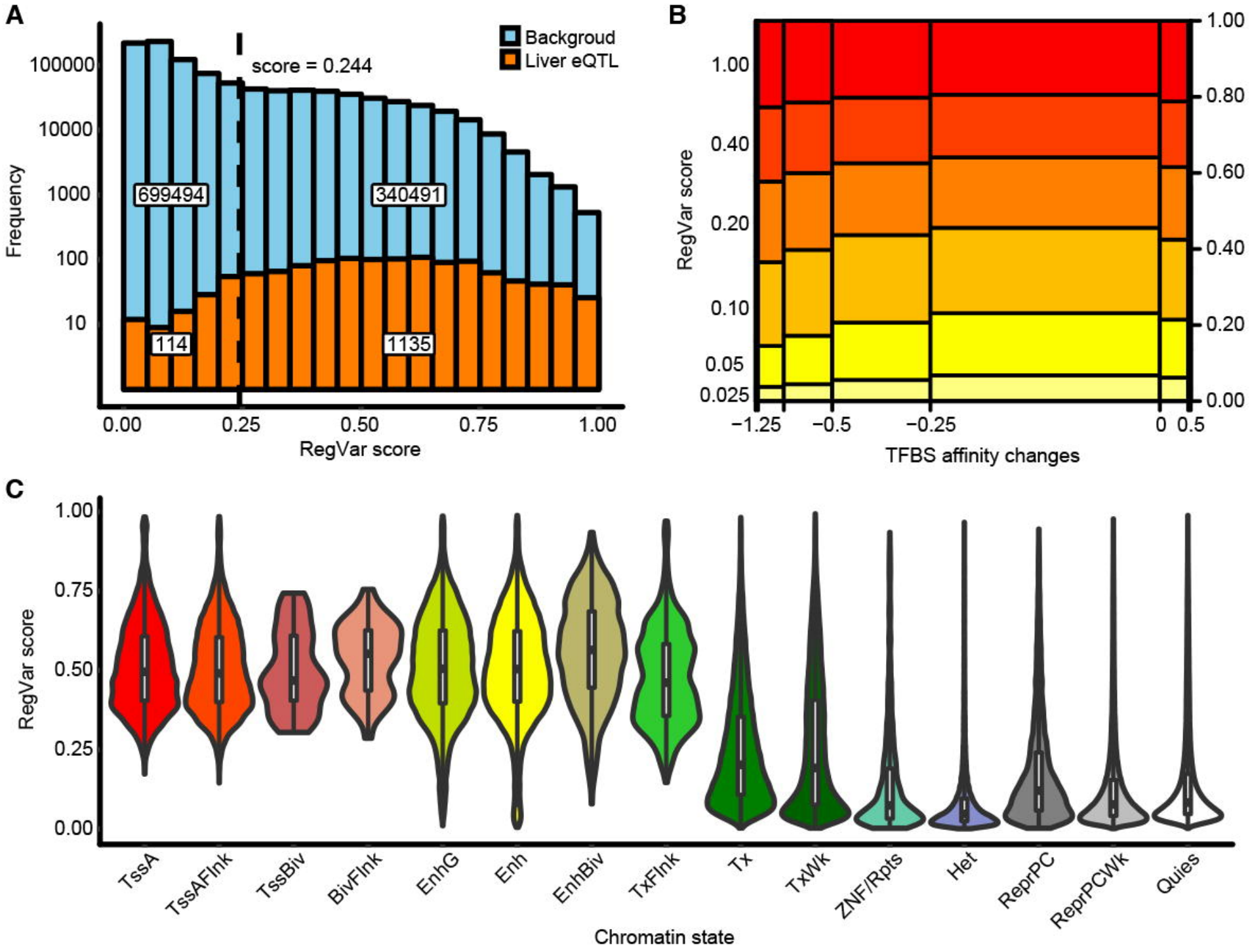
813 **File S1 Supplementary methods**

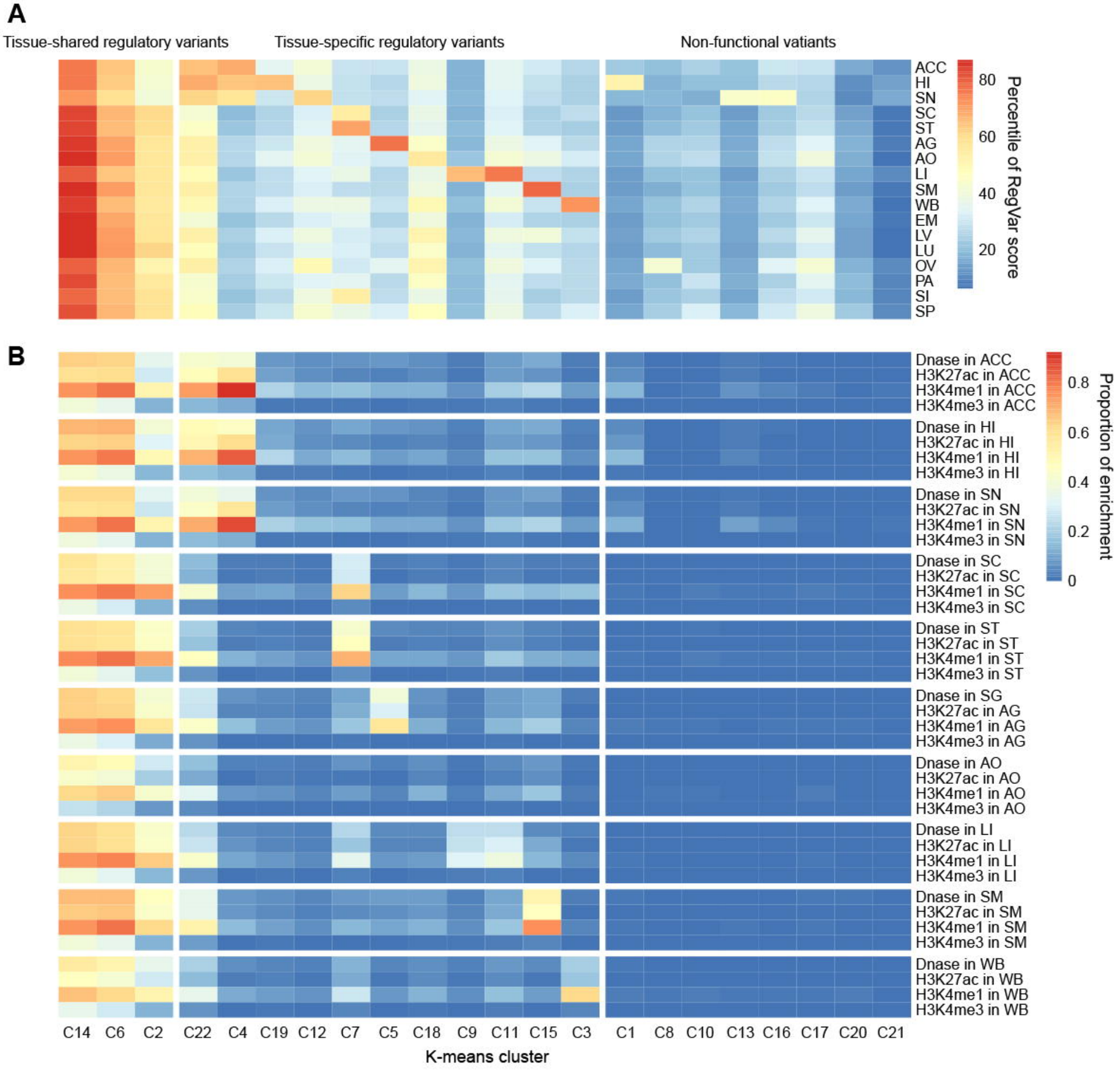
814

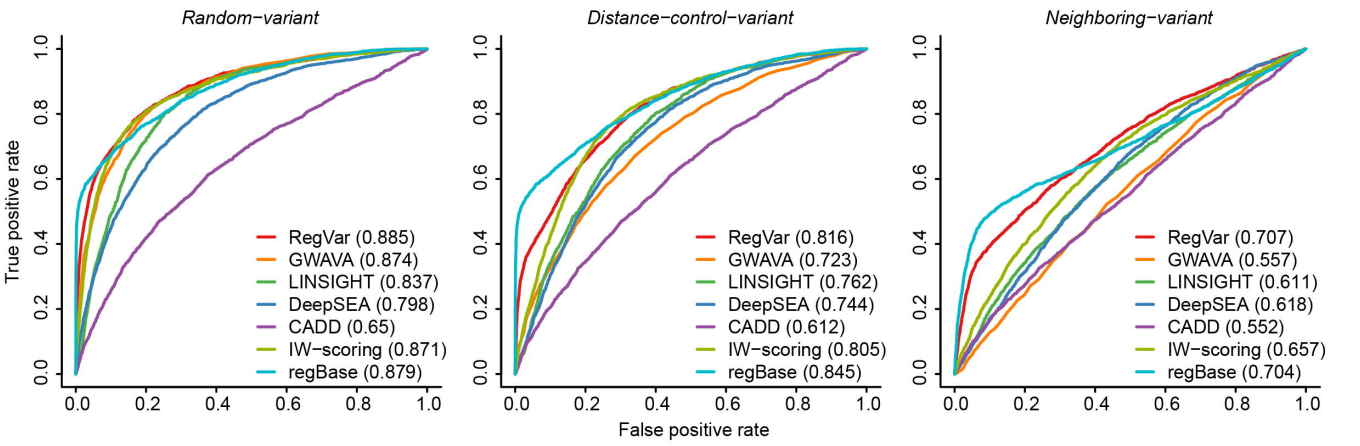

Dataset
Genomic profiles
Model
Positive dataset
GTEx SNV-TSS
Negative dataset
Random-variant set
Mirrored-variant set
Neighboring-variant set
Random-gene set
Feature mapping
Epigenome profiles
DHS
Histone modification
Chromatin state
FIRE
Sequential profiles
Sequential context
Genomic distance
TF binding affinity
Evolutionary profiles
PhyloP score
PhastCons score
Model training
DNN classifier
External evaluation
Prediction
SNV
Gene
RegVar score
rs375161066
NFIA
0.7604
rs117718808
FGD5
0.0103

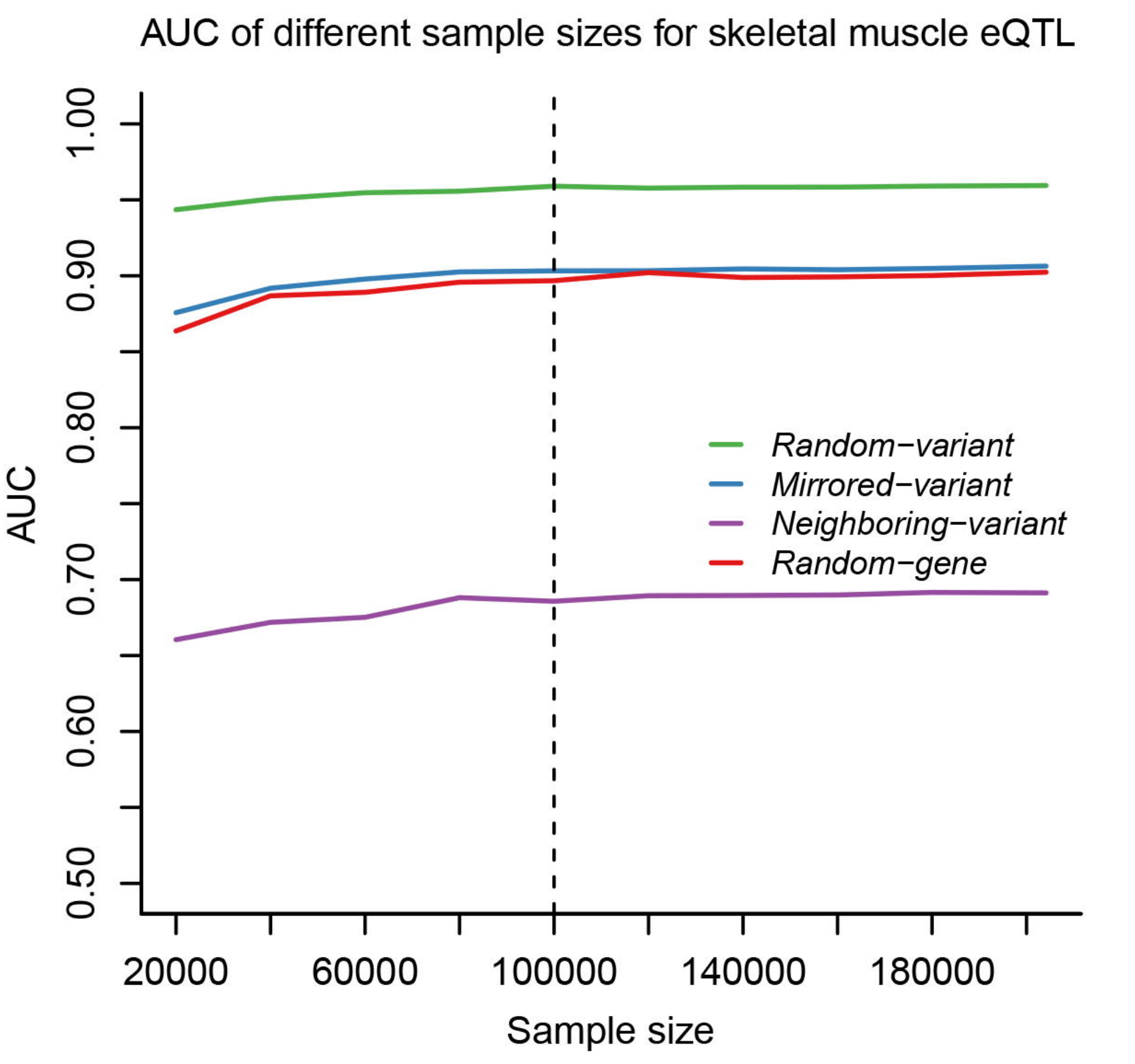

True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Random−variant (0.965)
Mirrored−variant (0.917)
Neighboring−variant (0.693)
Random−gene (0.929)

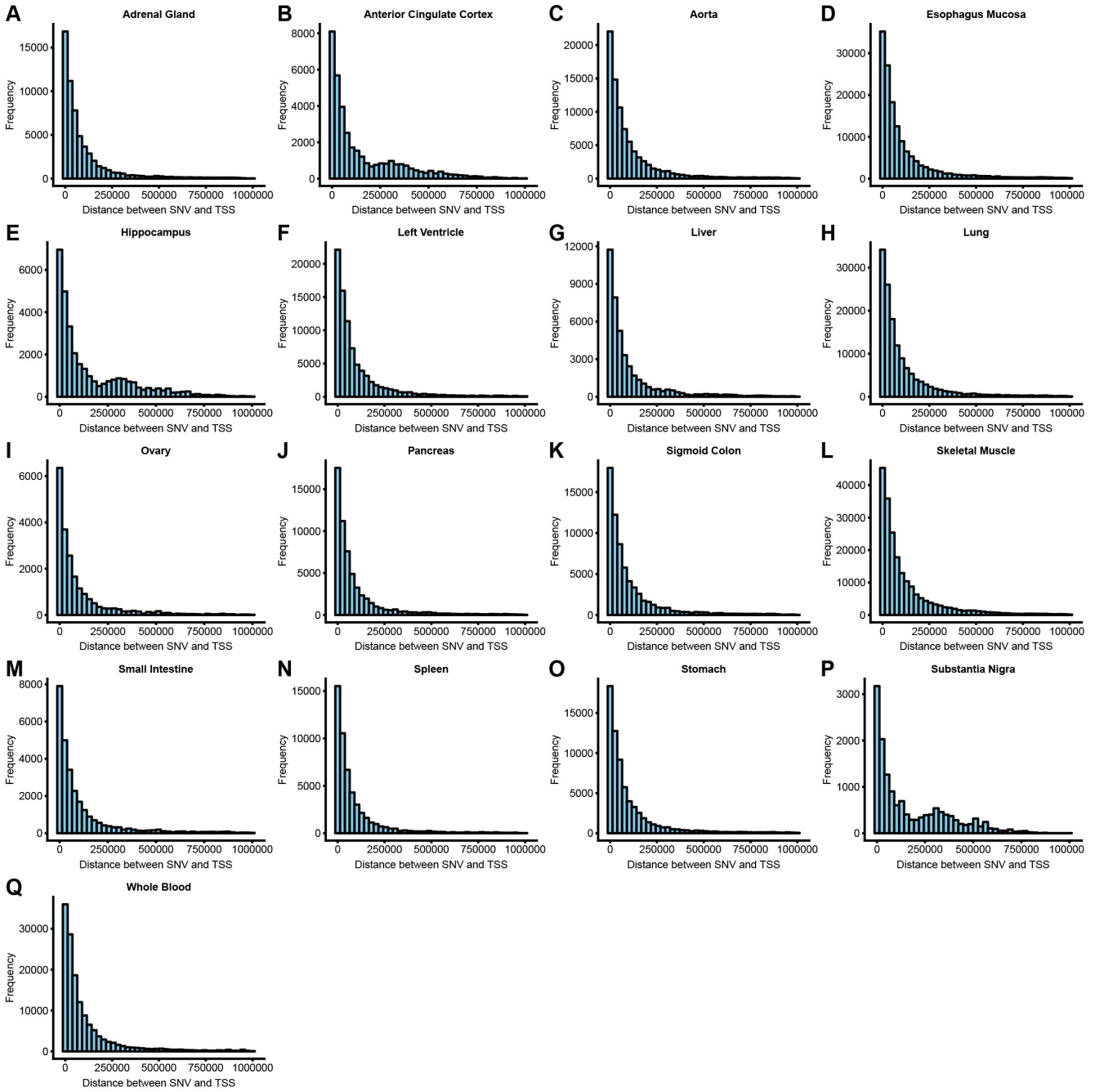

Random−variant
Mirrored−variant
Neighboring−variant
GTEx liver
exSNP liver
True positive rate
False positive rate
RegVar (0.969)
GWAVA (0.76)
LINSIGHT (0.717)
DeepSEA (0.57)
CADD (0.539)
TIVAN (0.653)
IW−scoring (0.671)
regBase (0.612)
RegVar (0.926)
GWAVA (0.627)
LINSIGHT (0.658)
DeepSEA (0.562)
CADD (0.521)
TIVAN (0.532)
ExPecto (0.54)
IW−scoring (0.593)
regBase (0.548)
RegVar (0.7)
GWAVA (0.488)
LINSIGHT (0.514)
DeepSEA (0.511)
CADD (0.467)
TIVAN (0.498)
ExPecto (0.525)
IW−scoring (0.485)
regBase (0.444)
RegVar (0.97)
GWAVA (0.759)
LINSIGHT (0.737)
DeepSEA (0.619)
CADD (0.536)
TIVAN (0.699)
IW−scoring (0.658)
regBase (0.619)
RegVar (0.893)
GWAVA (0.613)
LINSIGHT (0.674)
DeepSEA (0.611)
CADD (0.515)
TIVAN (0.541)
ExPecto (0.494)
IW−scoring (0.62)
regBase (0.543)
RegVar (0.724)
GWAVA (0.492)
LINSIGHT (0.522)
DeepSEA (0.541)
CADD (0.458)
TIVAN (0.495)
ExPecto (0.487)
IW−scoring (0.481)
regBase (0.445)

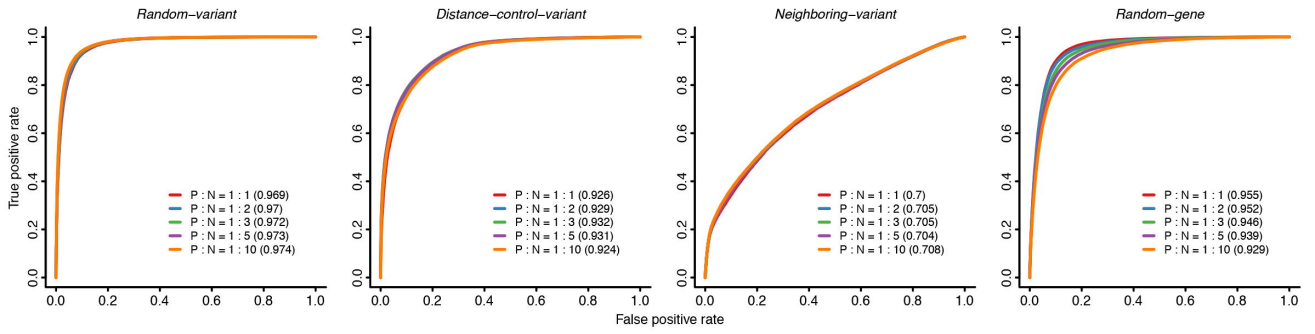

A
Backgroud
Liver eQTL
score = 0.244
699494
340491
114
1135
Frequency
100000
10000
1000
100
10
0.00
0.25
0.50
0.75
1.00
RegVar score
B
RegVar score
1.00
0.40
0.20
0.10
0.05
0.025
1.00
0.80
0.60
0.40
0.20
0.00
−1.25
−0.5
−0.25
0
0.5
TFBS affinity changes
C
RegVar score
1.00
0.75
0.50
0.25
0.00
TssA
TssAFlnk
TssBiv
BivFlnk
EnhG
Enh
EnhBiv
TxFlnk
Tx
TxWk
ZNF/Rpts
Het
ReprPC
ReprPCWk
Quies
Chromatin state

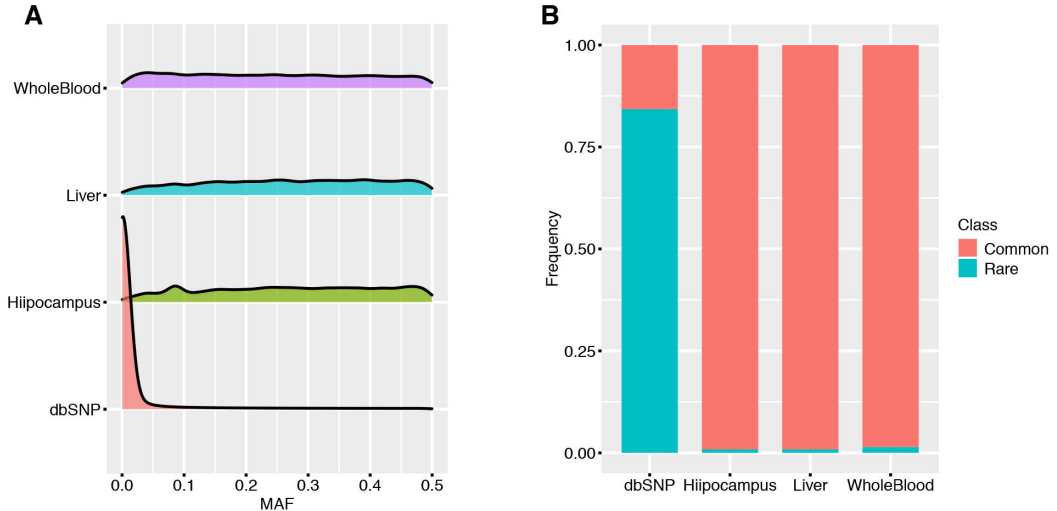

A
Tissue-shared regulatory variants
Tissue-specific regulatory variants
Non-functional vatiants
ACC
HI
SN
SC
ST
AG
AO
LI
SM
WB
EM
LV
LU
OV
PA
SI
SP
80
60
40
20
Percentile of RegVar score
B
Dnase in ACC
H3K27ac in ACC
H3K4me1 in ACC
H3K4me3 in ACC
Dnase in HI
H3K27ac in HI
H3K4me1 in HI
H3K4me3 in HI
Dnase in SN
H3K27ac in SN
H3K4me1 in SN
H3K4me3 in SN
Dnase in SC
H3K27ac in SC
H3K4me1 in SC
H3K4me3 in SC
Dnase in ST
H3K27ac in ST
H3K4me1 in ST
H3K4me3 in ST
Dnase in SG
H3K27ac in AG
H3K4me1 in AG
H3K4me3 in AG
Dnase in AO
H3K27ac in AO
H3K4me1 in AO
H3K4me3 in AO
Dnase in LI
H3K27ac in LI
H3K4me1 in LI
H3K4me3 in LI
Dnase in SM
H3K27ac in SM
H3K4me1 in SM
H3K4me3 in SM
Dnase in WB
H3K27ac in WB
H3K4me1 in WB
H3K4me3 in WB
0.8
0.6
0.4
0.2
0
Proportion of enrichment
C14
C6
C2
C22
C4
C19
C12
C7
C5
C18
C9
C11
C15
C3
C1
C8
C10
C13
C16
C17
C20
C21
K-means cluster

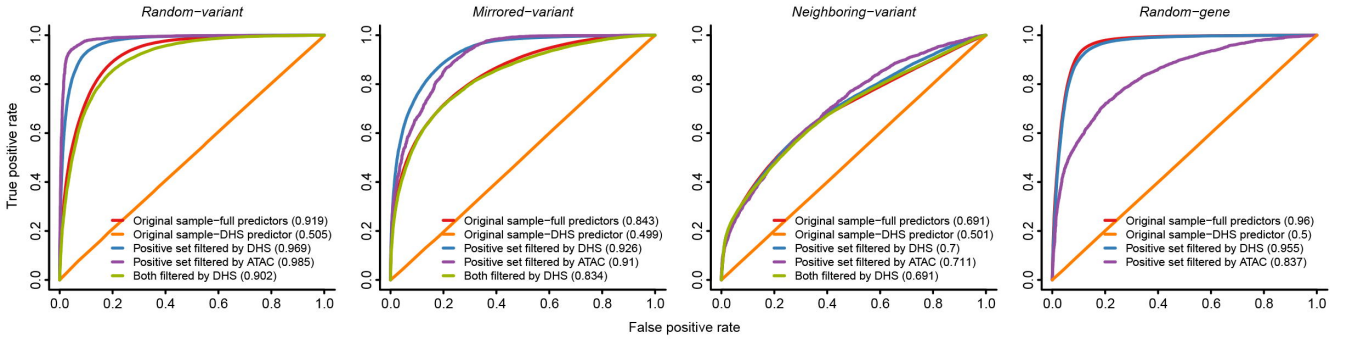

Random−variant
Distance−control−variant
Neighboring−variant
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RegVar (0.885)
GWAVA (0.874)
LINSIGHT (0.837)
DeepSEA (0.798)
CADD (0.65)
IW−scoring (0.871)
regBase (0.879)
RegVar (0.816)
GWAVA (0.723)
LINSIGHT (0.762)
DeepSEA (0.744)
CADD (0.612)
IW−scoring (0.805)
regBase (0.845)
RegVar (0.707)
GWAVA (0.557)
LINSIGHT (0.611)
DeepSEA (0.618)
CADD (0.552)
IW−scoring (0.657)
regBase (0.704)

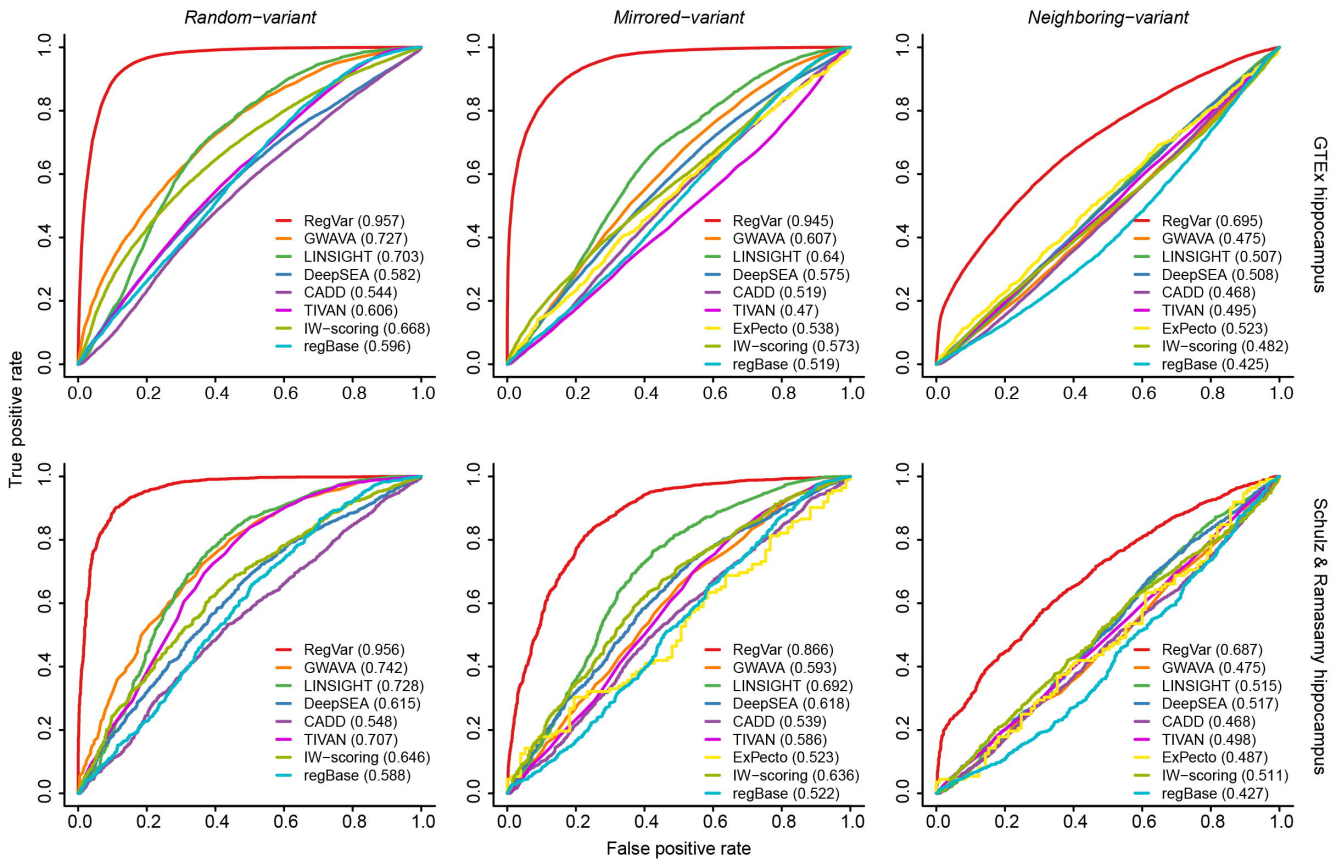

AUC of different sample sizes for skeletal muscle eQTL
AUC
1.00
0.90
0.80
0.70
0.60
0.50
20000
60000
100000
140000
180000
Sample size
Random−variant
Mirrored−variant
Neighboring−variant
Random−gene

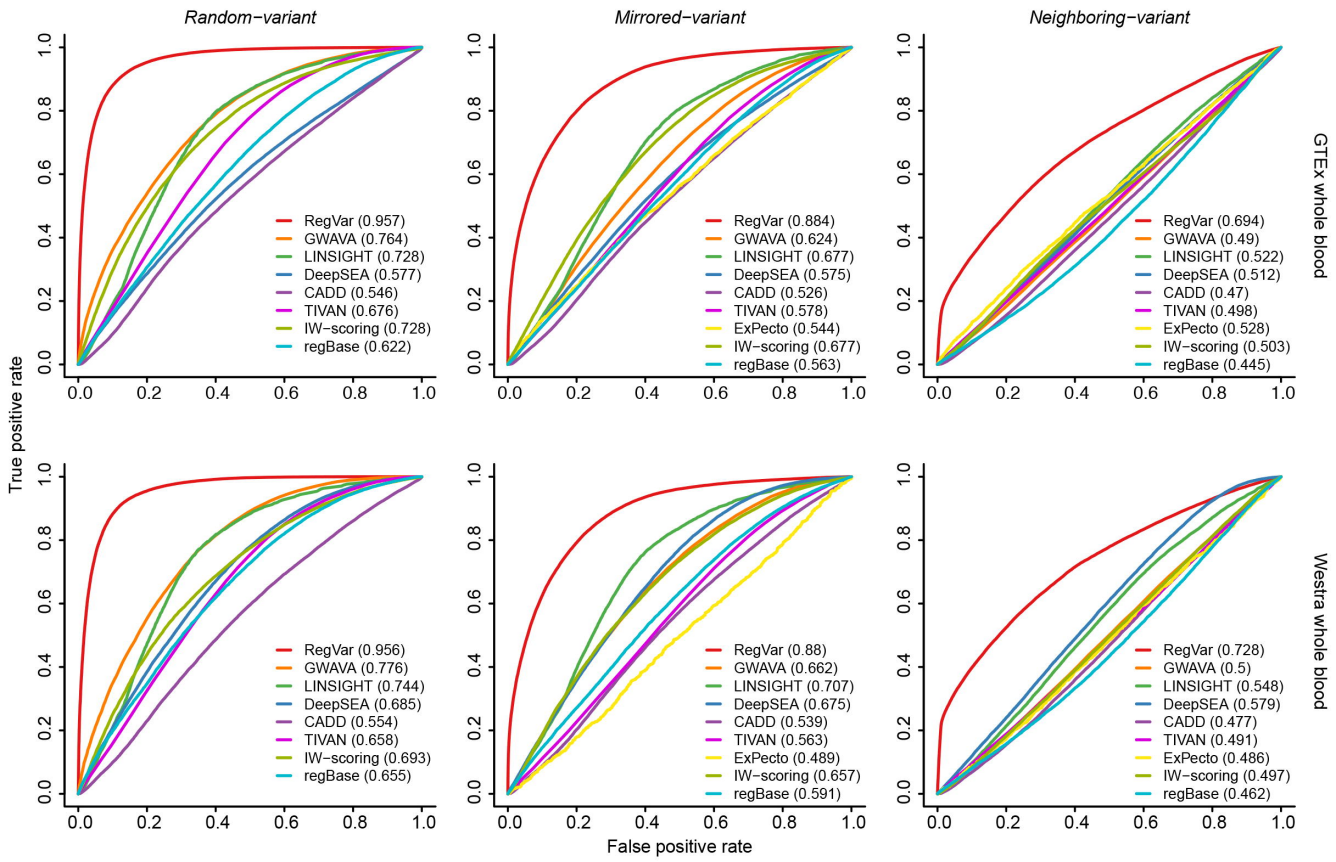

A
Adrenal Gland
B
Anterior Cingulate Cortex
C
Aorta
D
Esophagus Mucosa
E
Hippocampus
F
Left Ventricle
G
Liver
H
Lung
I
Ovary
J
Pancreas
K
Sigmoid Colon
L
Skeletal Muscle
M
Small Intestine
N
Spleen
O
Stomach
P
Substantia Nigra
Q
Whole Blood
Frequency
Distance between SNV and TSS

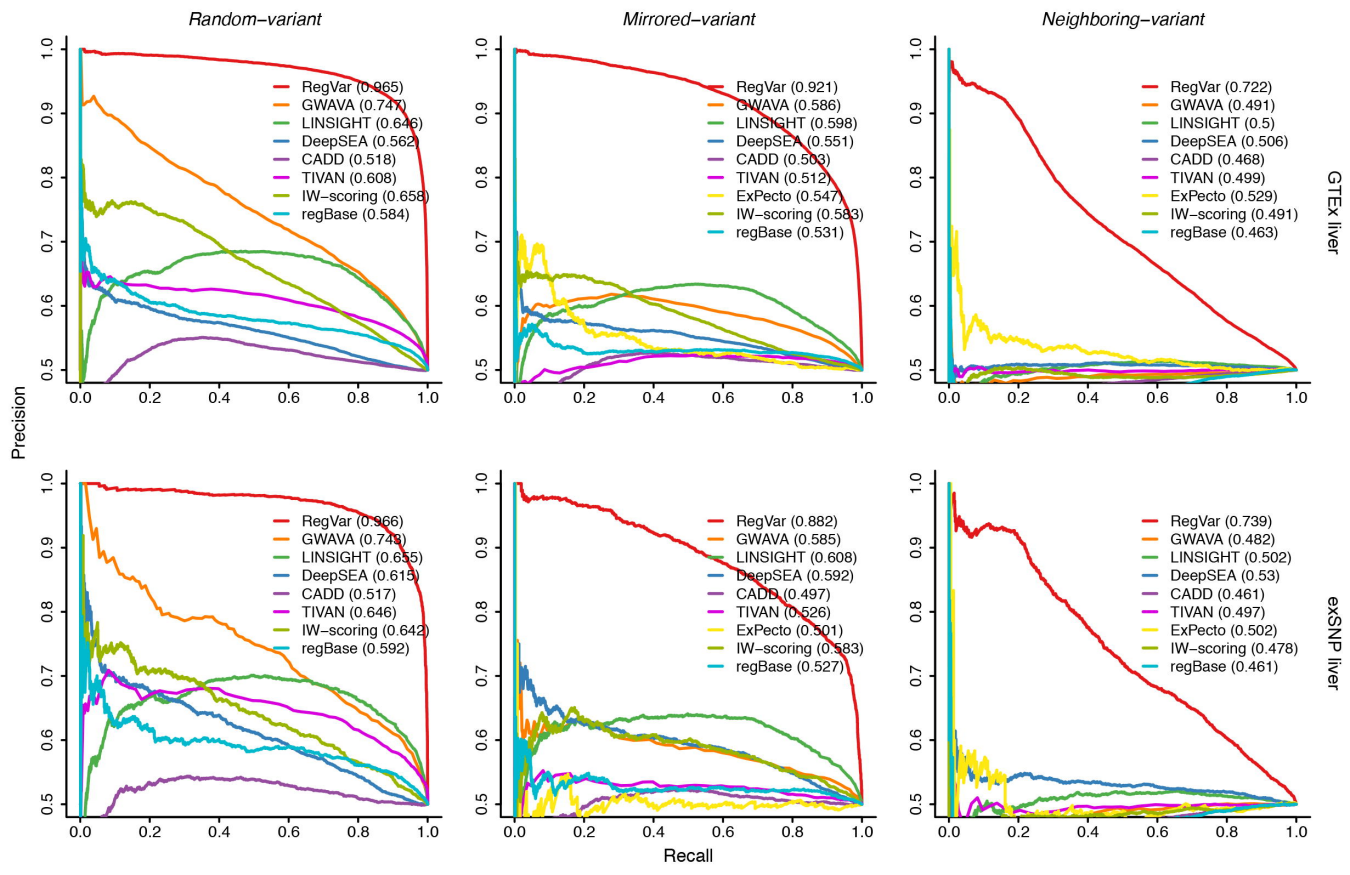

Random−variant
Distance−control−variant
Neighboring−variant
Random−gene
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
P : N = 1 : 1 (0.969)
P : N = 1 : 2 (0.97)
P : N = 1 : 3 (0.972)
P : N = 1 : 5 (0.973)
P : N = 1 : 10 (0.974)
P : N = 1 : 1 (0.926)
P : N = 1 : 2 (0.929)
P : N = 1 : 3 (0.932)
P : N = 1 : 5 (0.931)
P : N = 1 : 10 (0.924)
P : N = 1 : 1 (0.7)
P : N = 1 : 2 (0.705)
P : N = 1 : 3 (0.705)
P : N = 1 : 5 (0.704)
P : N = 1 : 10 (0.708)
P : N = 1 : 1 (0.955)
P : N = 1 : 2 (0.952)
P : N = 1 : 3 (0.946)
P : N = 1 : 5 (0.939)
P : N = 1 : 10 (0.929)

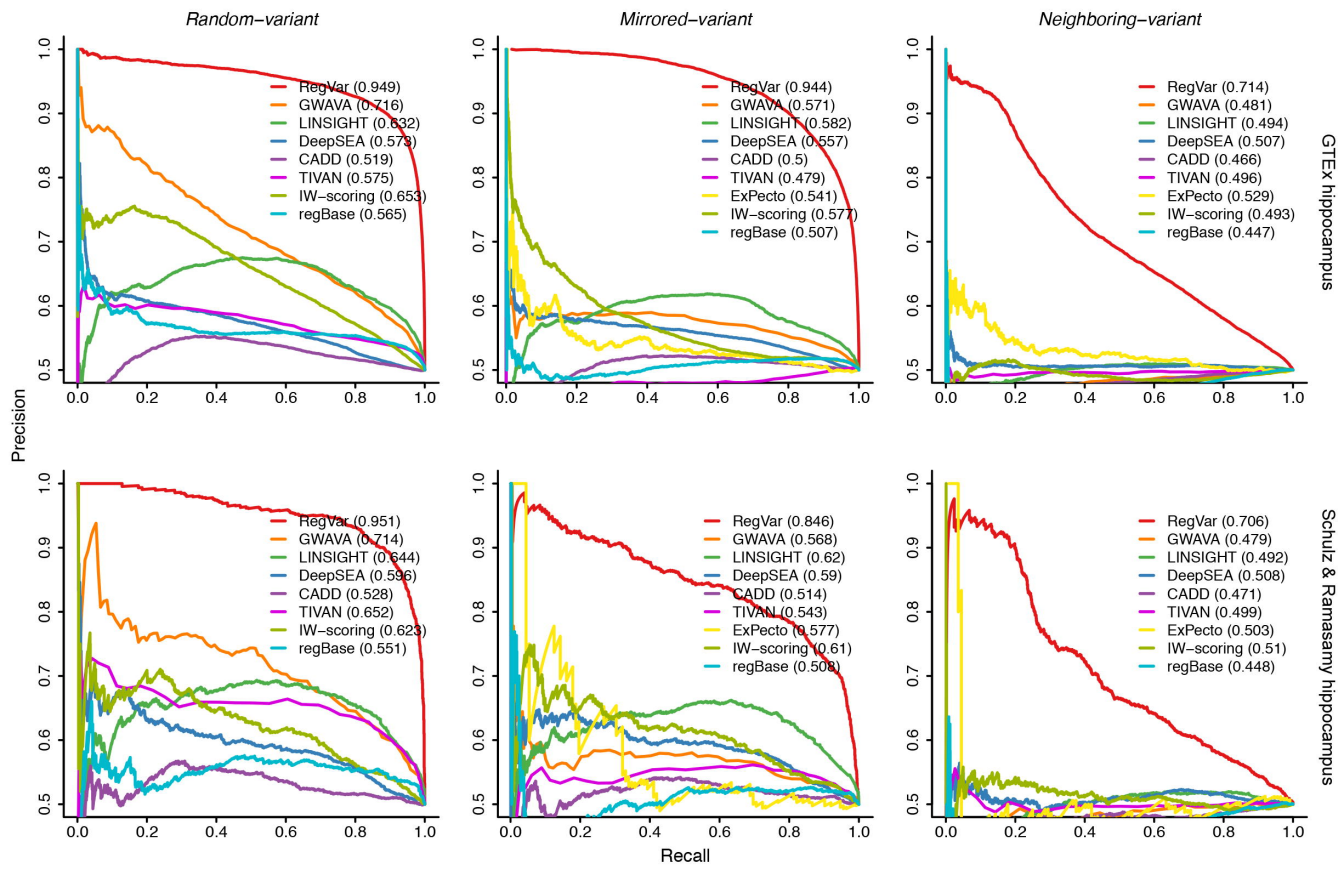

A
WholeBlood
Liver
Hiipocampus
dbSNP
0.0
0.1
0.2
0.3
0.4
0.5
MAF
B
1.00
0.75
0.50
0.25
0.00
Frequency
dbSNP
Hiipocampus
Liver
WholeBlood
Class
Common
Rare

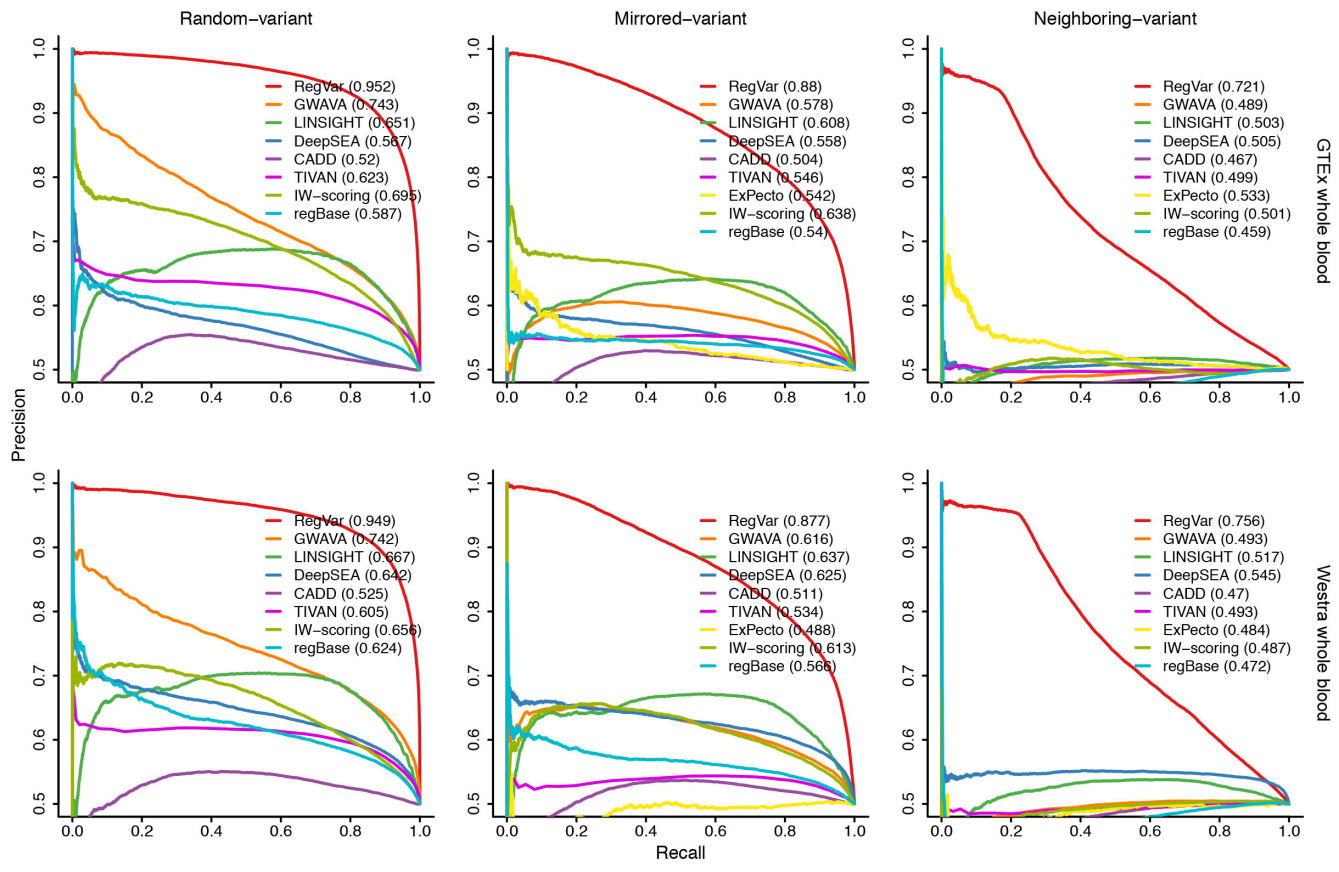

Random−variant
Original sample−full predictors (0.919)
Original sample−DHS predictor (0.505)
Positive set filtered by DHS (0.969)
Positive set filtered by ATAC (0.985)
Both filtered by DHS (0.902)
Mirrored−variant
Original sample−full predictors (0.843)
Original sample−DHS predictor (0.499)
Positive set filtered by DHS (0.926)
Positive set filtered by ATAC (0.91)
Both filtered by DHS (0.834)
Neighboring−variant
Original sample−full predictors (0.691)
Original sample−DHS predictor (0.501)
Positive set filtered by DHS (0.7)
Positive set filtered by ATAC (0.711)
Both filtered by DHS (0.691)
Random−gene
Original sample−full predictors (0.96)
Original sample−DHS predictor (0.5)
Positive set filtered by DHS (0.955)
Positive set filtered by ATAC (0.837)
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

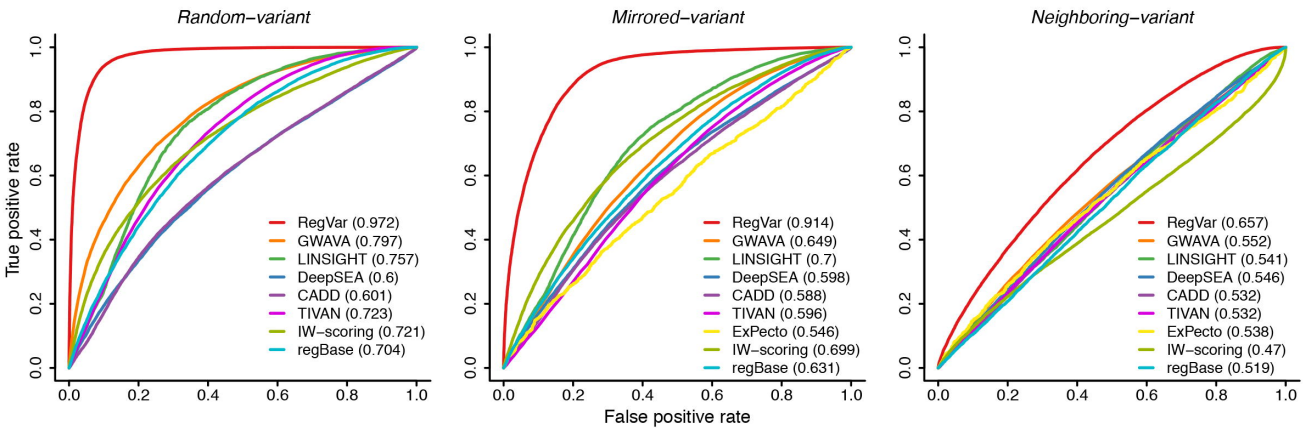

Random−variant
Mirrored−variant
Neighboring−variant
GTEx hippocampus
Schulz & Ramasamy hippocampus
True positive rate
False positive rate
RegVar (0.957)
GWAVA (0.727)
LINSIGHT (0.703)
DeepSEA (0.582)
CADD (0.544)
TIVAN (0.606)
IW−scoring (0.668)
regBase (0.596)
RegVar (0.945)
GWAVA (0.607)
LINSIGHT (0.64)
DeepSEA (0.575)
CADD (0.519)
TIVAN (0.47)
ExPecto (0.538)
IW−scoring (0.573)
regBase (0.519)
RegVar (0.695)
GWAVA (0.475)
LINSIGHT (0.507)
DeepSEA (0.508)
CADD (0.468)
TIVAN (0.495)
ExPecto (0.523)
IW−scoring (0.482)
regBase (0.425)
RegVar (0.956)
GWAVA (0.742)
LINSIGHT (0.728)
DeepSEA (0.615)
CADD (0.548)
TIVAN (0.707)
IW−scoring (0.646)
regBase (0.588)
RegVar (0.866)
GWAVA (0.593)
LINSIGHT (0.692)
DeepSEA (0.618)
CADD (0.539)
TIVAN (0.586)
ExPecto (0.523)
IW−scoring (0.636)
regBase (0.522)
RegVar (0.687)
GWAVA (0.475)
LINSIGHT (0.515)
DeepSEA (0.517)
CADD (0.468)
TIVAN (0.498)
ExPecto (0.487)
IW−scoring (0.511)
regBase (0.427)

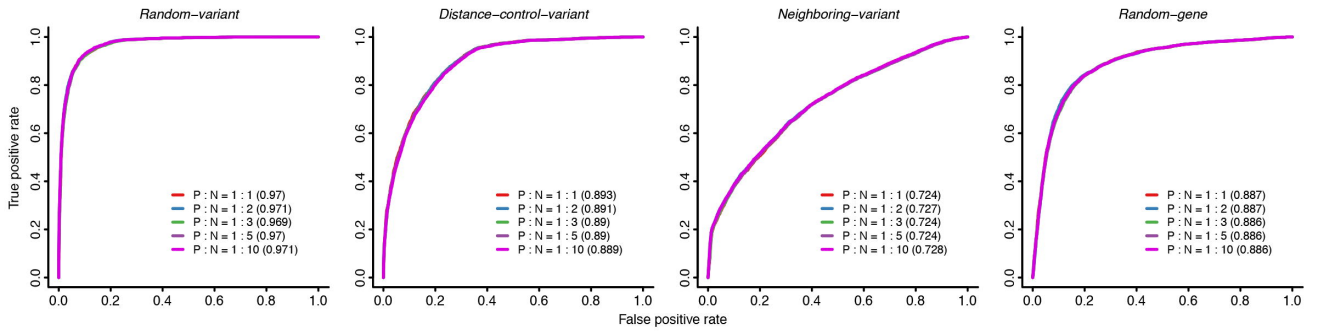

Random−variant
Mirrored−variant
Neighboring−variant
GTEx whole blood
Westra whole blood
True positive rate
False positive rate
RegVar (0.957)
GWAVA (0.764)
LINSIGHT (0.728)
DeepSEA (0.577)
CADD (0.546)
TIVAN (0.676)
IW−scoring (0.728)
regBase (0.622)
RegVar (0.884)
GWAVA (0.624)
LINSIGHT (0.677)
DeepSEA (0.575)
CADD (0.526)
TIVAN (0.578)
ExPecto (0.544)
IW−scoring (0.677)
regBase (0.563)
RegVar (0.694)
GWAVA (0.49)
LINSIGHT (0.522)
DeepSEA (0.512)
CADD (0.47)
TIVAN (0.498)
ExPecto (0.528)
IW−scoring (0.503)
regBase (0.445)
RegVar (0.956)
GWAVA (0.776)
LINSIGHT (0.744)
DeepSEA (0.685)
CADD (0.554)
TIVAN (0.658)
IW−scoring (0.693)
regBase (0.655)
RegVar (0.88)
GWAVA (0.662)
LINSIGHT (0.707)
DeepSEA (0.675)
CADD (0.539)
TIVAN (0.563)
ExPecto (0.489)
IW−scoring (0.657)
regBase (0.591)
RegVar (0.728)
GWAVA (0.5)
LINSIGHT (0.548)
DeepSEA (0.579)
CADD (0.477)
TIVAN (0.491)
ExPecto (0.486)
IW−scoring (0.497)
regBase (0.462)

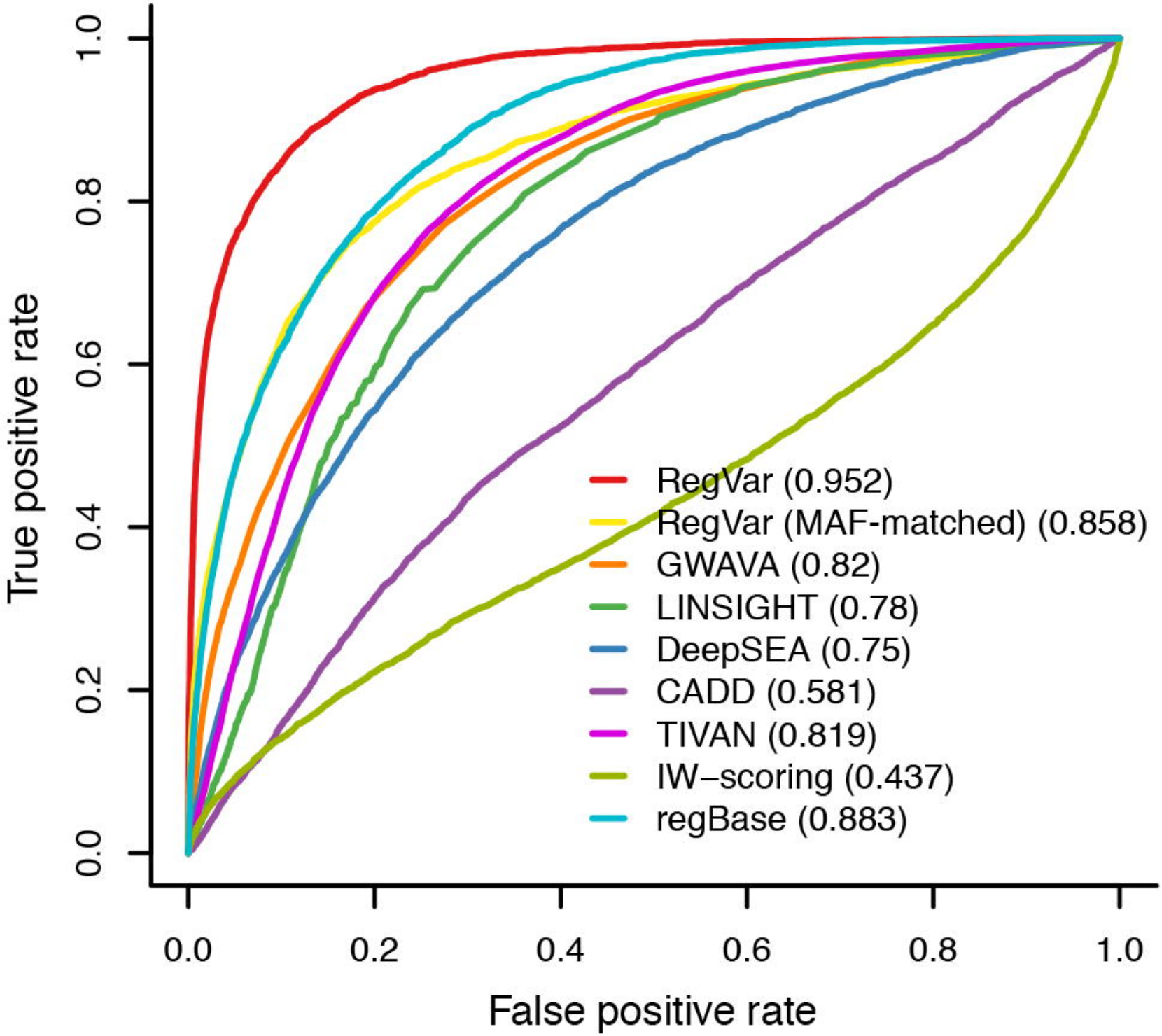

Random−variant
Mirrored−variant
Neighboring−variant
GTEx liver
exSNP liver
Precision
Recall
RegVar (0.965)
GWAVA (0.747)
LINSIGHT (0.646)
DeepSEA (0.562)
CADD (0.518)
TIVAN (0.608)
IW−scoring (0.658)
regBase (0.584)
RegVar (0.921)
GWAVA (0.586)
LINSIGHT (0.598)
DeepSEA (0.551)
CADD (0.503)
TIVAN (0.512)
ExPecto (0.547)
IW−scoring (0.583)
regBase (0.531)
RegVar (0.722)
GWAVA (0.491)
LINSIGHT (0.5)
DeepSEA (0.506)
CADD (0.468)
TIVAN (0.499)
ExPecto (0.529)
IW−scoring (0.491)
regBase (0.463)
RegVar (0.966)
GWAVA (0.748)
LINSIGHT (0.655)
DeepSEA (0.615)
CADD (0.517)
TIVAN (0.646)
IW−scoring (0.642)
regBase (0.592)
RegVar (0.882)
GWAVA (0.585)
LINSIGHT (0.608)
DeepSEA (0.592)
CADD (0.497)
TIVAN (0.526)
ExPecto (0.501)
IW−scoring (0.583)
regBase (0.527)
RegVar (0.739)
GWAVA (0.482)
LINSIGHT (0.502)
DeepSEA (0.53)
CADD (0.461)
TIVAN (0.497)
ExPecto (0.502)
IW−scoring (0.478)
regBase (0.461)

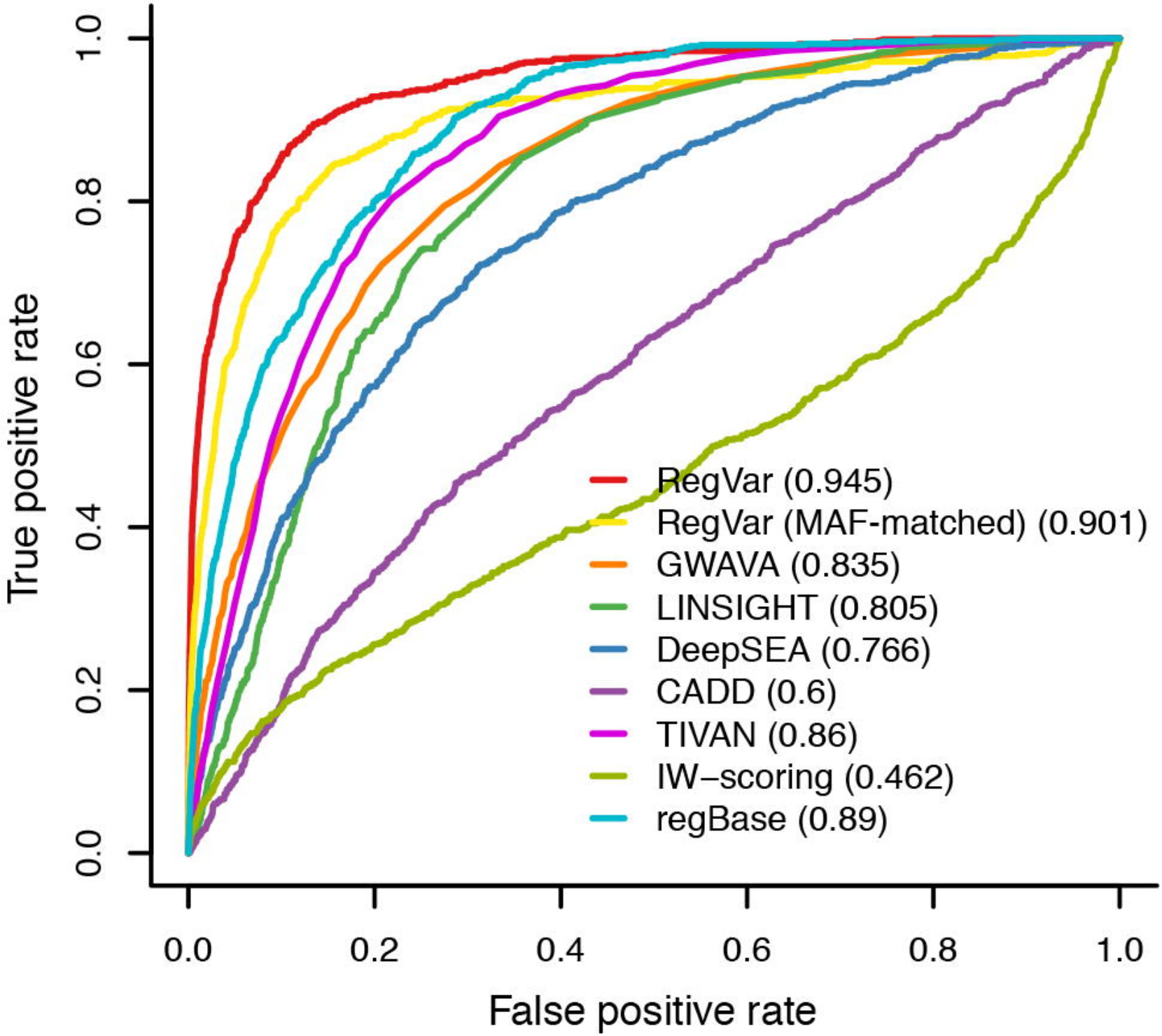

Random−variant
Mirrored−variant
Neighboring−variant
GTEx hippocampus
Schulz & Ramasamy hippocampus
Precision
Recall
RegVar (0.949)
GWAVA (0.716)
LINSIGHT (0.632)
DeepSEA (0.578)
CADD (0.519)
TIVAN (0.575)
IW−scoring (0.653)
regBase (0.565)
RegVar (0.944)
GWAVA (0.571)
LINSIGHT (0.582)
DeepSEA (0.557)
CADD (0.5)
TIVAN (0.479)
ExPecto (0.541)
IW−scoring (0.577)
regBase (0.507)
RegVar (0.714)
GWAVA (0.481)
LINSIGHT (0.494)
DeepSEA (0.507)
CADD (0.466)
TIVAN (0.496)
ExPecto (0.529)
IW−scoring (0.493)
regBase (0.447)
RegVar (0.951)
GWAVA (0.714)
LINSIGHT (0.644)
DeepSEA (0.596)
CADD (0.528)
TIVAN (0.652)
IW−scoring (0.623)
regBase (0.551)
RegVar (0.846)
GWAVA (0.568)
LINSIGHT (0.62)
DeepSEA (0.59)
CADD (0.514)
TIVAN (0.543)
ExPecto (0.577)
IW−scoring (0.61)
regBase (0.508)
RegVar (0.706)
GWAVA (0.479)
LINSIGHT (0.492)
DeepSEA (0.508)
CADD (0.471)
TIVAN (0.499)
ExPecto (0.503)
IW−scoring (0.51)
regBase (0.448)

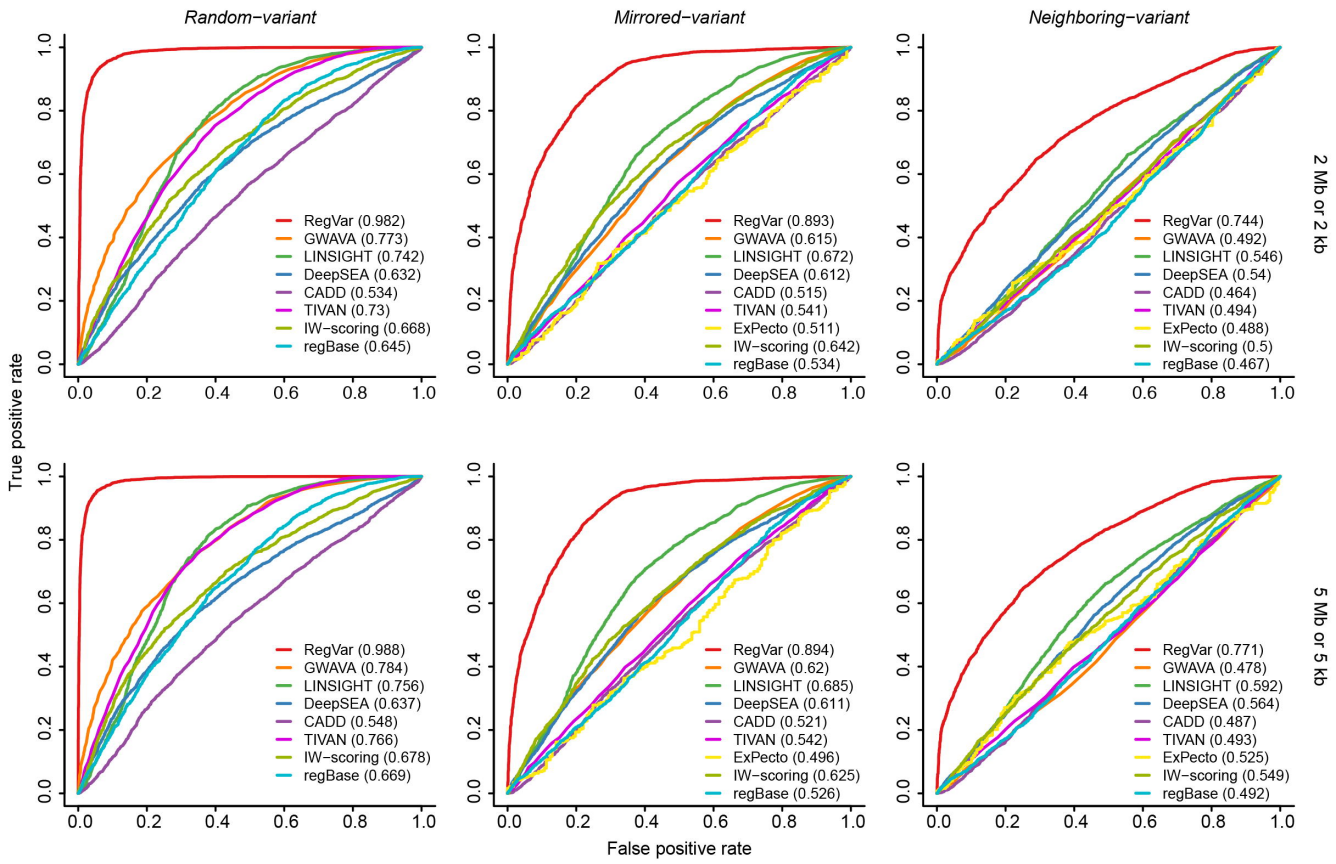

Random−variant
Mirrored−variant
Neighboring−variant
GTEx whole blood
Westra whole blood
Precision
Recall
RegVar (0.952)
GWAVA (0.743)
LINSIGHT (0.651)
DeepSEA (0.567)
CADD (0.52)
TIVAN (0.623)
IW−scoring (0.695)
regBase (0.587)
RegVar (0.88)
GWAVA (0.578)
LINSIGHT (0.608)
DeepSEA (0.558)
CADD (0.504)
TIVAN (0.546)
ExPecto (0.542)
IW−scoring (0.638)
regBase (0.54)
RegVar (0.721)
GWAVA (0.489)
LINSIGHT (0.503)
DeepSEA (0.505)
CADD (0.467)
TIVAN (0.499)
ExPecto (0.533)
IW−scoring (0.501)
regBase (0.459)
RegVar (0.949)
GWAVA (0.742)
LINSIGHT (0.667)
DeepSEA (0.642)
CADD (0.525)
TIVAN (0.605)
IW−scoring (0.656)
regBase (0.624)
RegVar (0.877)
GWAVA (0.616)
LINSIGHT (0.637)
DeepSEA (0.625)
CADD (0.511)
TIVAN (0.534)
ExPecto (0.488)
IW−scoring (0.613)
regBase (0.566)
RegVar (0.756)
GWAVA (0.493)
LINSIGHT (0.517)
DeepSEA (0.545)
CADD (0.47)
TIVAN (0.493)
ExPecto (0.484)
IW−scoring (0.487)
regBase (0.472)

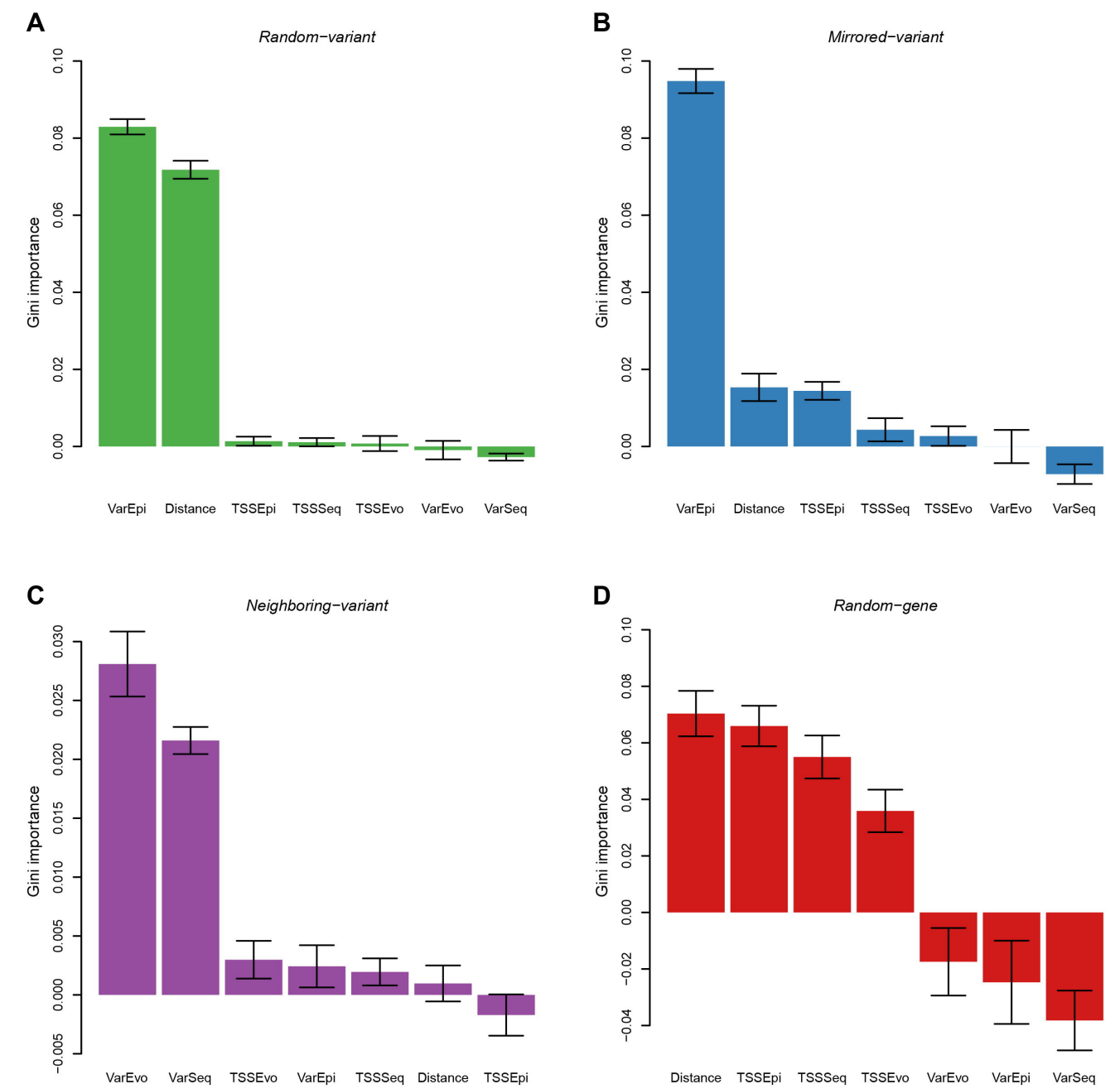

Random−variant
Mirrored−variant
Neighboring−variant
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RegVar (0.972)
GWAVA (0.797)
LINSIGHT (0.757)
DeepSEA (0.6)
CADD (0.601)
TIVAN (0.723)
IW−scoring (0.721)
regBase (0.704)
RegVar (0.914)
GWAVA (0.649)
LINSIGHT (0.7)
DeepSEA (0.598)
CADD (0.588)
TIVAN (0.596)
ExPecto (0.546)
IW−scoring (0.699)
regBase (0.631)
RegVar (0.657)
GWAVA (0.552)
LINSIGHT (0.541)
DeepSEA (0.546)
CADD (0.532)
TIVAN (0.532)
ExPecto (0.538)
IW−scoring (0.47)
regBase (0.519)

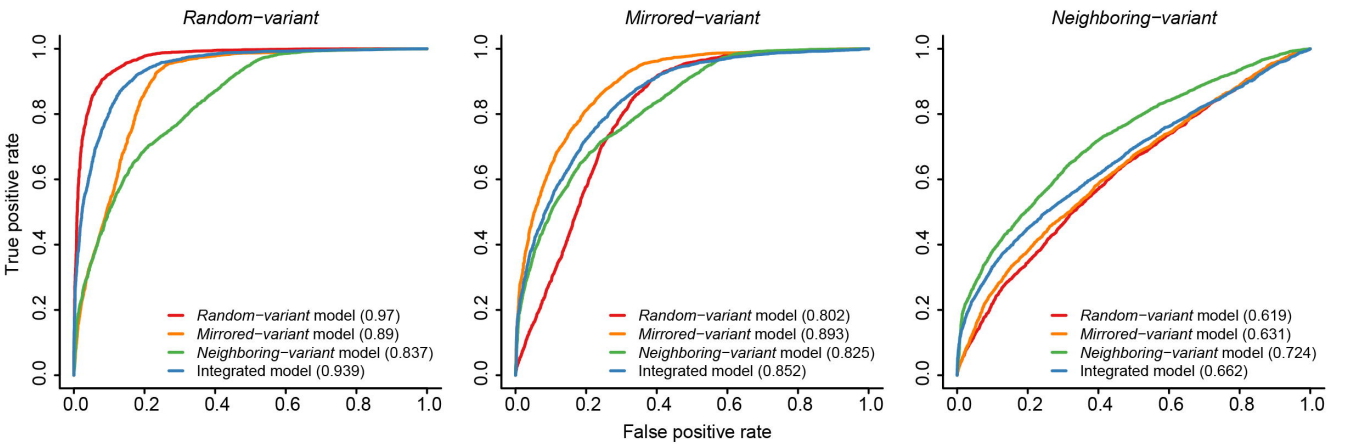

Random−variant
Distance−control−variant
Neighboring−variant
Random−gene
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
P : N = 1 : 1 (0.97)
P : N = 1 : 2 (0.971)
P : N = 1 : 3 (0.969)
P : N = 1 : 5 (0.97)
P : N = 1 : 10 (0.971)
P : N = 1 : 1 (0.893)
P : N = 1 : 2 (0.891)
P : N = 1 : 3 (0.89)
P : N = 1 : 5 (0.89)
P : N = 1 : 10 (0.889)
P : N = 1 : 1 (0.724)
P : N = 1 : 2 (0.727)
P : N = 1 : 3 (0.724)
P : N = 1 : 5 (0.724)
P : N = 1 : 10 (0.728)
P : N = 1 : 1 (0.887)
P : N = 1 : 2 (0.887)
P : N = 1 : 3 (0.886)
P : N = 1 : 5 (0.886)
P : N = 1 : 10 (0.886)

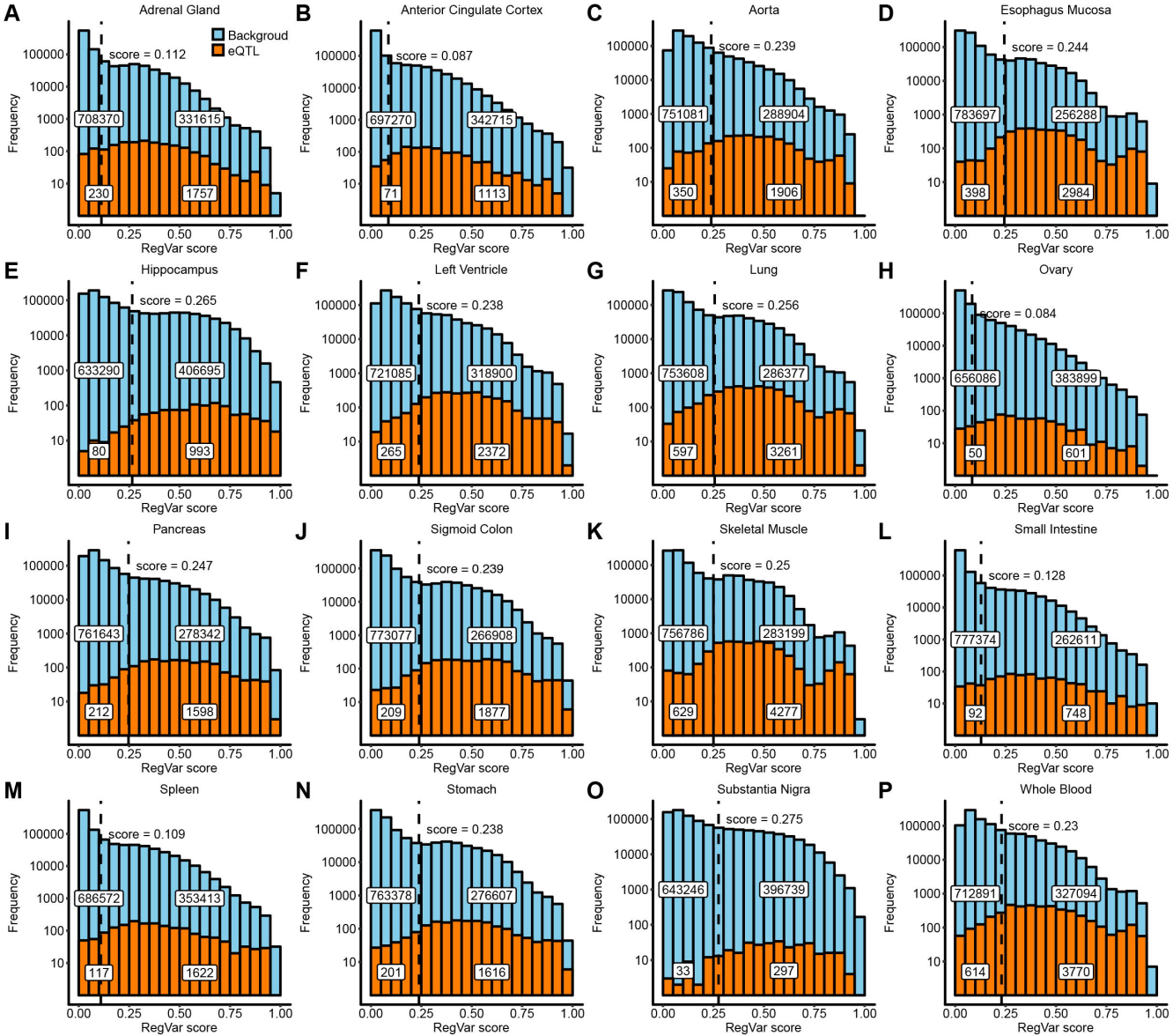

True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RegVar (0.952)
RegVar (MAF-matched) (0.858)
GWAVA (0.82)
LINSIGHT (0.78)
DeepSEA (0.75)
CADD (0.581)
TIVAN (0.819)
IW−scoring (0.437)
regBase (0.883)

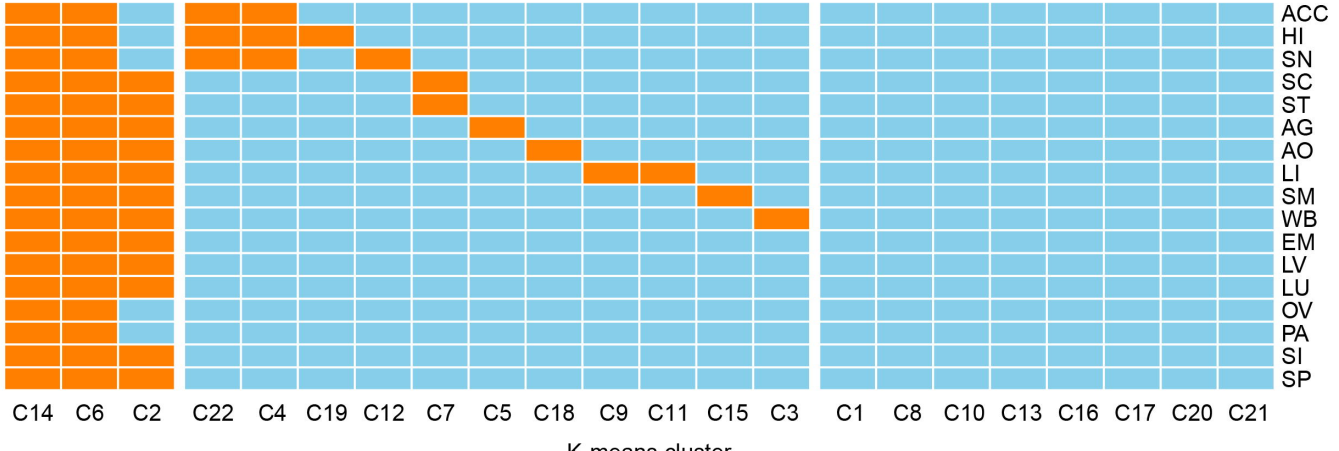

True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RegVar (0.945)
RegVar (MAF-matched) (0.901)
GWAVA (0.835)
LINSIGHT (0.805)
DeepSEA (0.766)
CADD (0.6)
TIVAN (0.86)
IW−scoring (0.462)
regBase (0.89)

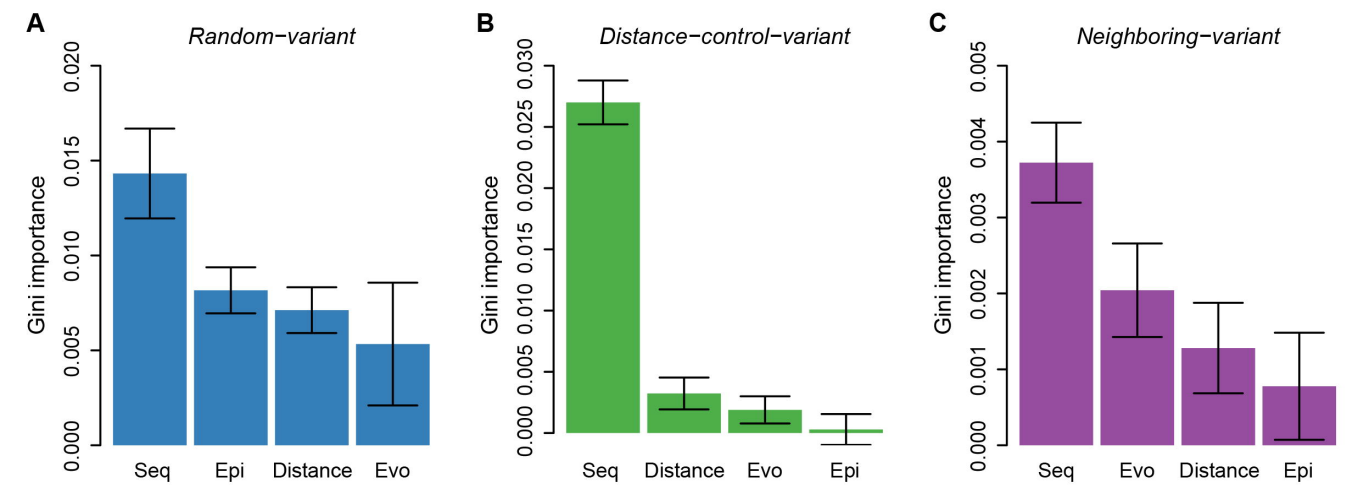

Random−variant
Mirrored−variant
Neighboring−variant
2 Mb or 2 kb
5 Mb or 5 kb
True positive rate
False positive rate
RegVar (0.982)
GWAVA (0.773)
LINSIGHT (0.742)
DeepSEA (0.632)
CADD (0.534)
TIVAN (0.73)
IW−scoring (0.668)
regBase (0.645)
RegVar (0.893)
GWAVA (0.615)
LINSIGHT (0.672)
DeepSEA (0.612)
CADD (0.515)
TIVAN (0.541)
ExPecto (0.511)
IW−scoring (0.642)
regBase (0.534)
RegVar (0.744)
GWAVA (0.492)
LINSIGHT (0.546)
DeepSEA (0.54)
CADD (0.464)
TIVAN (0.494)
ExPecto (0.488)
IW−scoring (0.5)
regBase (0.467)
RegVar (0.988)
GWAVA (0.784)
LINSIGHT (0.756)
DeepSEA (0.637)
CADD (0.548)
TIVAN (0.766)
IW−scoring (0.678)
regBase (0.669)
RegVar (0.894)
GWAVA (0.62)
LINSIGHT (0.685)
DeepSEA (0.611)
CADD (0.521)
TIVAN (0.542)
ExPecto (0.496)
IW−scoring (0.625)
regBase (0.526)
RegVar (0.771)
GWAVA (0.478)
LINSIGHT (0.592)
DeepSEA (0.564)
CADD (0.487)
TIVAN (0.493)
ExPecto (0.525)
IW−scoring (0.549)
regBase (0.492)


A
Random−variant
Gini importance
0.00
0.02
0.04
0.06
0.08
0.10
VarEpi
Distance
TSSEpi
TSSSeq
TSSEvo
VarEvo
VarSeq
B
Mirrored−variant
Gini importance
0.00
0.02
0.04
0.06
0.08
0.10
VarEpi
Distance
TSSEpi
TSSSeq
TSSEvo
VarEvo
VarSeq
C
Neighboring−variant
Gini importance
−0.005
0.000
0.005
0.010
0.015
0.020
0.025
0.030
VarEvo
VarSeq
TSSEvo
VarEpi
TSSSeq
Distance
TSSEpi
D
Random−gene
Gini importance
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
Distance
TSSEpi
TSSSeq
TSSEvo
VarEvo
VarEpi
VarSeq


Random−variant
Mirrored−variant
Neighboring−variant
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Random−variant model (0.97)
Mirrored−variant model (0.89)
Neighboring−variant model (0.837)
Integrated model (0.939)
Random−variant model (0.802)
Mirrored−variant model (0.893)
Neighboring−variant model (0.825)
Integrated model (0.852)
Random−variant model (0.619)
Mirrored−variant model (0.631)
Neighboring−variant model (0.724)
Integrated model (0.662)


A
Adrenal Gland
Backgroud
eQTL
score = 0.112
708370
331615
230
1757
B
Anterior Cingulate Cortex
score = 0.087
697270
342715
71
1113
C
Aorta
score = 0.239
751081
288904
350
1906
D
Esophagus Mucosa
score = 0.244
783697
256288
398
2984
E
Hippocampus
score = 0.265
633290
406695
80
993
F
Left Ventricle
score = 0.238
721085
318900
265
2372
G
Lung
score = 0.256
753608
286377
597
3261
H
Ovary
score = 0.084
656086
383899
50
601
I
Pancreas
score = 0.247
761643
278342
212
1598
J
Sigmoid Colon
score = 0.239
773077
266908
209
1877
K
Skeletal Muscle
score = 0.25
756786
283199
629
4277
L
Small Intestine
score = 0.128
777374
262611
92
748
M
Spleen
score = 0.109
686572
353413
117
1622
N
Stomach
score = 0.238
763378
276607
201
1616
O
Substantia Nigra
score = 0.275
643246
396739
33
297
P
Whole Blood
score = 0.23
712891
327094
614
3770
Frequency
RegVar score
0.00
0.25
0.50
0.75
1.00
10
100
1000
10000
100000


ACC
HI
SN
SC
ST
AG
AO
LI
SM
WB
EM
LV
LU
OV
PA
SI
SP
C14 C6 C2 C22 C4 C19 C12 C7 C5 C18 C9 C11 C15 C3 C1 C8 C10 C13 C16 C17 C20 C21
K-means cluster


A
Random−variant
Gini importance
0.000
0.005
0.010
0.015
0.020
Seq
Epi
Distance
Evo
B
Distance−control−variant
Gini importance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Seq
Distance
Evo
Epi
C
Neighboring−variant
Gini importance
0.000
0.001
0.002
0.003
0.004
0.005
Seq
Evo
Distance
Epi