# TVAR: Assessing Tissue-specific Functional Effects of Non-coding Variants with Deep Learning

**Hai Yang**
East China University of Science and Technology
**Rui Chen**
Vanderbilt University
**Quan Wang**
Vanderbilt University
**Qiang Wei**
Vanderbilt University
**Ying Ji**
Vanderbilt University https://orcid.org/0000-0001-5691-1303
**Xue Zhong**
Vanderbilt University
**Bingshan Li** ( bingshan.li@Vanderbilt.Edu )
Vanderbilt University https://orcid.org/0000-0003-2129-168X

**Article**

**Keywords:** whole genome-sequencing, TVAR, deep learning, non-coding variants

**Posted Date:** December 2nd, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-113771/v1

# TVAR: Assessing Tissue-specific Functional Effects of Non-coding Variants with Deep Learning

Hai Yang[1,2,3], Rui Chen[2,3], Quan Wang[2,3], Qiang Wei[2,3], Ying Ji[2,3], Xue Zhong[3,4], Bingshan Li [2,3,*]

1. Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

2. Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, Tennessee, United States of America

3. Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, United States of America

4. Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America.

* Corresponding author

Correspondence should be sent to: Bingshan Li, bingshan.li@vanderbilt.edu

**Abstract**

Analysis of whole genome-sequencing (WGS) for genetics of disease is still a challenge due to lack of accurate functional annotation of noncoding variants, especially the rare ones. As eQTLs have been extensively implicated in genetics of human diseases, we hypothesize that noncoding rare variants discovered in WGS play a regulatory role in predisposing disease risk. With thousands of tissue- and cell type-specific epigenomic features, we propose TVAR, a multi-label learning based deep neural network that predicts the functionality of noncoding variants in the genome based on eQTLs across 49 human tissues in GTEx. TVAR learns the relationships between high-dimensional epigenomics and eQTLs across tissues, taking the correlation among tissues into account to learn shared and tissue-specific eQTL effects. As a result, TVAR outputs tissue-specific annotations, with an average of 0.77 across these tissues. We evaluate TVAR's performance on four complex diseases (coronary artery disease, breast cancer, Type 2 diabetes, and Schizophrenia), using TVAR's tissue-specific annotations, and observe its superior performance in predicting functional variants for both common and rare variants, compared to five existing state-of-the-art tools. We further evaluate TVAR's G-score, a scoring scheme across all tissues, on ClinVar, fine-mapped GWAS loci, Massive Parallel Reporter Assay (MPRA) validated variants, and observe consistently better performance of TVAR compared to other competing tools.

## Introduction

With the rapid development of Whole-genome sequencing (WGS), the interpretation of non-coding regions accounting for 98% of the human genome has become a significant focus of genetic studies[1]. Studies have shown that the majority of non-coding variants associated with complex diseases predispose disease risk by regulating the expression of their target genes, often in a tissue-specific manner[2]. Understanding the regulatory role of the non-coding variants will help advance research in genetics, evolution, and precision medicine[3]. Genome-wide association studies (GWAS) have identified thousands of disease-associated variants, most of which are common in population[4]. In contrast, analysis of rare noncoding variants suffers from low power and severe multiple testing correction. Functional annotation of noncoding variants becomes essential to prioritize potentially risk variants for complex diseases. Massive genomics data, e.g. ENCODE[5], Roadmap[6] and FANTOM5[7] have been used in various computational methods (CADD[8], GWAVA[9], DANN[10], LINSIGHT[11], EIGEN[12], DVAR [13]) for such purposes. These methods are generally based on machine learning, supervised or unsupervised. In general, supervised learning is powerful when equipped with well-calibrated training set; however, for annotating noncoding variants, these methods are often troubled by the lack of high-accuracy training set[12]. On the other hand, unsupervised methods e.g. Eigen[12] , PINES[14] and DVAR[13], discover the internal differences among the groups of variants in the training set, without the need of labeling of true functional variants, and are therefore more robust in the situation without curated functional variants that are representative of risk variants for complex

disease. It is well established that risk variants for complex diseases have weak effect sizes, even for rare variants[15]. Known disease causing noncoding rare variants in Clinvar for example are usually biased towards variants with large effect sizes, making the machine learning approaches based on such training set unlikely to accurately annotate functional variants for complex diseases. Strategies that are able to model representative functional variants for complex diseases are needed.

Risk variants from GWAS exert disease risk largely through regulating risk genes. Expression QTLs (eQTLs) have been shown to significantly co-localize with GWAS loci and used to identify risk genes regulated by risk variants[16]. Thus, eQTLs provide a robust means to functionally annotate risk variants for complex diseases. An added layer of complexity is that both risk variants and gene expression exhibit strong tissue-specificity[17]. It is therefore unlikely that a single annotation is equally effective for different diseases. We argue that strategies taking tissue specificity of both the functional variants and target genes are effective in prioritizing functional variants for complex diseases. Large-scale tissue and cell type specific epigenomics data, e.g. ENCODE[5], Roadmap Epigenomics[18], have been used in various machine learning methods to infer functionality of noncoding variants. Transcriptome and matched WGS data across a wide range of tissues in GTEx provide genome-wide tissue-specific eQTLs. The reported eQTLs provide robust and unbiased association signals that provide opportunities for accurately identifying tissue-specific risk variants.

For a long time, due to the limitation of the size of the existing annotation

database and the computer power, most computational models (such as the above approaches) only describe the functional effects of variants at the organicism level. Until recently, a wide range of tissue /cell type-specific functional annotations (such as ENCODE, Roadmap, and GTEx[19]) are published, and computational methods begin to estimate the issue-specific functionality of variants. Due to the paucity of known noncoding variants associated with complex diseases, it is challenging to train machine learning approaches using well-validated labels. For this reason, very few methods, e.g. Delta-SVM[20] and GenoNet[21], use supervised training for the prediction of tissue-specific functional effects of noncoding variants. Other methods, e.g. FUN-LDA[22], FitCons2[4], and ExPecto[23], employed unsupervised learning approaches that do not rely on accurately labeled data on various tissues/cell types. These methods, however, train models individually for each tissue and do not consider the correlation between tissues, resulting in loss of accuracy as well as potential incompatibility among tissues. The challenge of estimating the functionality of variants in various tissues is far from being considered resolved.

The following issues need to be addressed urgently to address the challenge of predicting tissue-specific non-coding functional variants. As most eQTLs detectable with sufficient power are common, how to leverage the eQTL resources to predict functionality of rare noncoding variants is a key challenge and opportunity. An additional challenge is how to leverage the genetic sharing and dissimilarity of eQTLs across tissues to achieve tissue specific prediction. To address the first challenge, our rationale is that common and rare regulatory variants share similar regulatory

machinery such that they exist similar functional genomics data across tissues and cell types. Based on the rationale, we develop the prediction model to link genomics data to eQTLs, and apply the model to predict functionality of rare variants based on the shared genomics data. For the second challenge, we need to consider the connections between multiple tissues simultaneously, and more specifically need to build a joint model of each eQTL across multiple tissues, which requires a reasonable function that is able to map multi-dimensional functional genomics data to multi-tissue eQTLs.

In this study, we propose TVAR, a deep neural network (DNN) that integrates multi-label learning and multi-instance learning to solve the above problems. By using the 1247-dimensional functional genomics features, TVAR accesses the tissue-specific functional scores of each variant across the 49 GTEx tissues (from the GTEx dataset V8). Based on the multi-label DNN, during the training, TVAR learns the differences and connections between tissues, and jointly considers the functional utility of a variant across 49 tissues simultaneously to leverage the sharing of eQTL among tissues. Meanwhile, we also developed a multi-instance learning algorithm, G-score, to provide an integrated functional score for each variant on the organism level. We observed that TVAR achieves an average AUC of ROC of 0.77 on more than 70,000 variant datasets in 49 tissues of the GTEx dataset, and found that although the functional input annotations are the same, the maximum and minimum performances of TVAR across various tissues differ significantly, indicating that TVAR captures tissue-specific regulatory machinery in high-dimensional genomics data to make tissue-specific functional prediction of regulatory variants. We applied TVAR scoring on GWAS

data of 4 complex diseases, i.e. coronary artery disease, breast cancer, Type 2 diabetes, and Schizophrenia, with tissue-specific annotations from heart, breast, pancreas, and brains tissues in GTEx. We found that the top scoring rare variants (MAF<0.01) have significantly smaller *P* values than the background variants for each of the diseases, supporting that TVAR can be used to prioritize noncoding rare variants in a tissue-specific manner for complex diseases. Finally, we compared the performance of the TVAR G-score algorithm with other organism level score algorithms in four test scenarios (Clinvar[24,25] set, fine-mapped GWAS[26] set, GTEX eQTLs[27] set, and MPRA validated variants[28] set), and found that TVAR has superior performance across all the test scenarios. In particular, TVAR outperformed DeepSEA and DANN, two methods that are also based on deep learning methods. Overall, we demonstrate that the integration of multi-label learning and multi-instance learning with DNN implemented in TVAR holds promise for scoring non-coding variants in a tissue-specific manner for genetic analysis of complex diseases.

## RESULTS

### Evaluation of TVAR's performance across 49 tissues in the GTEx dataset

Through the TVAR framework, we can predict the functional variants corresponding to the 49 different human tissues in the GTEx (V8 release) (See Fig. 1 for the list of the tissues). Since TVAR is a supervised learning approach, to achieve the goal of identifying tissue-specific functional variants, sufficient and high-quality labels of training individuals are required. To obtain the training labels, we first collected the

variants that strongly associated with the eGenes in the GTEx dataset (with q-value cutoff 0.01). Due to the effect of the linkage disequilibrium (LD), most detected eQTLs are proxies to the functional regulatory variants [29]. We used a fine mapping strategy to nominate the credible functional variants in an eQTL LD block. Specifically, we used LINSIGHT, a state-of-art functional variant annotation tool, to filter these non-functional variants (Methods). For each tissue, we retained the top 1500 variants with the highest LINSIGHT scores, and merged the variant labels in 49 tissues into a matrix, in which each row is a variant, and each column is a tissue. The label matrix is a 0-1 matrix, which is used for the multi-label DNN. Since the variants in the label matrix show functionality in at least one tissue, we set these variants as the positive sample set for model training. Next, we randomly selected variants with MAF> 0.05 in the 1000 Genomes Project [30] to match the number of positive sample sets and removed the variants that appeared in the positive set to form the negative sample set. The labels of the negative sample set are all-zeros. Finally, we combined the matrix of the positive and negative samples to obtain a highly reliable training data set (70232 variants) derived from the GTEx cohort. We used q-value cutoff ($q < 0.05$) in all tissues to screen positive samples, and observed that out of the 35116 variants, 10421 (~30%) appeared in at least two tissues and 3455 (~10%) appeared in at least four tissues. We consider this strategy of accurately learning and successfully describing the associations between variants and tissues will be the passkey to predicting tissue-specific functional variants.

The TVAR framework is a deep feedforward network based on multi-label learning. The network of TVAR describes the functionality of the variant-tissue pairs through the fully connected layers, which can learn the differences and similarities among the 49 tissues. The output of TVAR is a 49-dimensional vector that represents the functional scores of the variant-tissue pairs. To prove that TVAR can successfully learn the functionality of the variants corresponding to each issue, we used five-fold cross-validation to train and test the model. In each training process, we randomly selected 80% of the data for the model training and 20% of the data for the model testing (see Supplementary Note 1). Receiver Operating Characteristic (ROC) curves were adopted to distinguish the prediction power of different methods in all testing processes. As an overall evaluation, we combined all variants across all tissues as a single evaluation set, and obtained an average AUC = 0.770 across the GTEx 49 tissues, indicating that through multi-label learning, the network can extract valid features from the input 1247-dimensional annotations predictive of eQTLs. Based on the average accuracy rate, we divided the GTEx 49 tissues into the 'high accuracy group' and the 'low accuracy group' (Fig. 1A).

In the high accuracy group (Fig. 1A), The AUCs of 15 tissues exceed 0.80, including 'Nerve_Tibial' (AUC=0.854), 'Cells_Transformed_fibroblasts' (AUC=0.851), 'Artery_Tibial' (AUC = 0.846) , 'Skin_Sun_Exposed_Lower_leg' (AUC = 0.843), 'Thyroid' (AUC = 0.841), 'Adipose_Subcutaneous' (AUC = 0.836), 'Muscle_Skeletal '(AUC = 0.834), 'Whole_Blood’ (0.826), 'Esophagus_Muscularis' (AUC = 0.826), 'Skin_Not_Sun_Exposed_Suprapubic’ (AUC = 0.825), 'Esophagus_Mucosa' (AUC =

0.822), 'Lung' (AUC = 0.814), 'Testis' (AUC = 0.808), ' Artery_Aorta' (AUC = 0.802), and ‘Adipose_VisceralO' (AUC = 0.802). On the remaining 10 tissues, the AUCs of TVAR exceeded (or approached) the average accuracy rate (AUC = 0.77): 'Colon_Transverse' (AUC = 0.795), 'Pancreas' (AUC = 0.792), 'Heart_Atrial_Appendage' (AUC = 0.790), 'Breast_Mammary_Ttissue' (AUC = 0.787), 'Esophagus_Gastroesophageal_Junction' (AUC = 0.785), 'Heart_Left_Ventricle (AUC = 0.784)', 'Colon_Sigmoid' (AUC = 0.781), 'Stomach' (AUC = 0.779), 'Sple’ (AUC = 0.777), 'Brain_Cerebellum' (AUC = 0.768). We found that TVAR often achieved similar AUCs on similar tissues, such as 'Skin_Sun_Exposed_Lower_leg' and ‘Skin_Not_Sun_Exposed_Suprapubic', 'Adipose_Subcutaneous' and 'Adipose_Visceral_Omentum', 'Heart_Left_Ventricle' and 'Heart_Atrial_Appendage',' Colon_Transverse ' and' Colon_Sigmiod'.

In the low performance group (Fig. 1B), The AUCs of 80% tissues exceed 0.70, including 'Pituitary' (AUC = 0.767), 'Adrenal_Gland' (AUC = 0.762), 'Pituitary' (AUC=0.752), 'Brain_Cerebellar_Hemisphere' (AUC=0.751), 'Brain_Cortex' (AUC=0.748), 'Brain_Nucleus_accumbens_basal_ganglia' (AUC=0.747), 'Brain_Caudate_basal_ganglia' (AUC=0.745), 'Brain_Frontal_Cortex_BA9' (AUC=0.742), 'Cells_EBV-transformed_lymphocytes' (AUC=0.742), 'Small_Intestine_Terminal_Ileum' (AUC=0.733), 'Artery_Coronary' (AUC=0.733), 'Brain_Putamen_basal_ganglia' (AUC=0.732), 'Liver' (AUC=0.724), ’Ovary’ (AUC=0.724), 'Brain_Hypothalamus' (AUC=0.722), 'Uterus' (AUC=0.719), 'Brain_Hippocampus' (AUC=0.719), 'Vagina' (AUC=0.713), 'Brain_Anterior_cingulate_cortex_BA24' (AUC=0.712), 'Minor_Salivary_Gland' (AUC=0.709). TVAR's performance on the remaining 4 tissues

in the low accuracy group is as follows: 'Brain_Amygdala' (AUC = 0.699), 'Brain_Substantia_nigra' (AUC = 0.696), ‘Kidney_Cortex' ( AUC = 0.695), 'Brain_Spinal_cord_cervical_c-1' (AUC = 0.690). TVAR's predictions in related tissues (such as brain tissues) are also similar (Table S2).

After the above analysis, we found that TVAR achieved the best performance on 'Nerve_Tibial' (AUC = 0.854), and the lowest performance on 'Brain_Substantia_nigra' (AUC = 0.690). The vast differences in performance indicate that the input features of the model have different interpretation capabilities for functional variants in different tissues. Overall, TVAR achieved reasonable predict power (AUC> 0.7) on the majority of the tissues (more than 91%).

**The relevance of TVAR scores across the 49 GTEx tissues**

After the training process with more than 70,000 variants on the GTEx dataset, TVAR successfully achieved stable performance in predicting functional variants on the GTEx 49 tissues. We aim to investigate the predictive scores across the tissues. Overall, it is clear that the more unrelated the tissues are, the higher dissimilarity of the TVAR scores are. To explore this phenomenon further, we selected the most significant eQTLs (q-value <1e-30) separately in each of the 49 tissues of the GTEx dataset (we removed the variants in the training set of the TVAR), and formed a testing dataset (with 1240 variants); the variants in this dataset can be considered the most tissue-specific functional variants. Next, we use the TVAR framework to score all variants in

this dataset. For each tissue, the TVAR outputs a 1240-dimensional scoring vector. We use these scoring vectors to calculate the distance matrix of 49 tissues and use agglomerative hierarchical clustering to analyze the patterns of TVAR scores across different tissues. Similar to the results of cis-eQTLs across the GTEx tissues[19], TVAR's scoring pattern confers the specificity of tissues (Fig. 2). For example, most tissues related to the brain are adjacent and located in the upper half of the figure, and two heart-related tissues are adjacent (Fig. 2). Other tissues have similar clustering, e.g. two colon related tissues, two adipose related tissues, and two esophagus related tissues (Fig. 2). It should be emphasized that the tissue-specific patterns in these results are solely derived from the knowledge learned by the TVAR model during the previous training process. The clustering patterns suggest that the TVAR scoring results are indeed tissue-specific, and the same variant will get similar scores in related tissues, and vice versa.

**Identification of the tissue-specific functionality of GWAS Hits**

To further investigate the usefulness of TVAR in predicting disease risk variants, we selected four complex diseases, i.e. coronary artery disease (CAD), breast cancer (BRC), Type 2 diabetes (T2D), and Schizophrenia (SCZ), and mapped the diseases to relevant tissues ('Heart_Left_Ventricle' for CAD, 'Breast_Mammary_Tissue' for BRC, 'Pancreas' for T2D, and 'Brain_Hippocampus' for SCZ). We used the TVAR scoring algorithm to predict GWAS hits on these diseases. We collected GWAS hits from the GWAS catalog (All associations v1.0) with $P<5*10^{-8}$, removing variants used in the TVAR training set.

In total, we obtained 325 GWAS hits for CAD, 345 GWAS hits for BRC, 200 GWAS hits for T2D, and 285 GWAS hits for SCZ. We randomly selected an equal number of variants from 1000 Genomes Project [31] (MAF> 0.05) as the negative samples for each disease. To eliminate the influence of LD on the positive and negative sample sets, we set the distance between all the variants in the negative sample set and the variants in the positive sample set above 100kb. Previous studies have shown that the functional scores of the positive sample set should be significantly higher than that of randomly selected variants[12]. We obtained the TVAR scores corresponding to these variants in the four tissues related to CAD, BRC, T2D, and SCZ.

We compared TVAR's performance with five other state-of-art methods: CADD, Eigen, DANN, DeepSEA, and FUN-LDA. CADD is a supervised learning method based on SVM, which is a widely used variant score evaluation tool. DANN uses a deep neural network to replace the SVM module in CADD, and performs better than CADD in certain scenarios[10]; Eigen is a general variant evaluation method based on unsupervised learning, and DeepSEA is a functional variant evaluation method based on deep convolutional networks. FUN-LDA is a recent work that is based on unsupervised learning and promotes tissue-specific functional variants prediction (it is also the only method we investigated that supports scoring on a large number of tissues and the corresponding scores on these tissues are available online). For CADD, Eigen, DANN, and DeepSEA, we directly took the score corresponding to the variant for the comparison. For FUN-LDA, it is based on the RoadMap 111 tissues, which are different from the 49 tissues used in TVAR. Here we chose E095 ('Left ventricle') related

to CAD, E028 ('Breast vHMEC mammary epithelial') related to BRC, E098 ('Pancreas') related to T2D, and E069 ('Brain cingulate gyrus') related to SCZ as the corresponding tissues of FUN-LDA on the four diseases, and calculated the FUN-LDA scores of the related variants.

We used one-sided Wilcoxon signed rank test to test whether each algorithm's scores in the positive sample set (e.g. GWAS hits) are significantly higher than the scores in the negative sample set. We observed that TVAR, along with four other methods (CADD, Eigen, DANN, and DeepSEA), all achieved significance in distinguishing positive and negative variants on all data sets (Fig. 3, Table 1). The exception is DANN, which achieved significance on GWAS data of all diseases except SCZ (Fig. 3, Table 1). As a direct comparison, TVAR outperformed all other methods in all of the diseases investigated here. For example, for CAD, TVAR achieved the best performance (*P* = 1.21e-24), followed by FUN-LDA (*P* = 3.25e-11), DeepSEA (*P* = 5.29e-10), Eigen (*P* = 1.93e-4), DANN (*P* = 1.14E-2), CADD (*P* = 2.30e-2); for BRC, TVAR also obtained the best performance (P = 1.60e-21), followed by DeepSEA (*P* = 8.20e-8), Eigen (*P* = 2.02e-7), DANN (*P* = 6.20e-6), CADD (*P* = 5.52e-5), FUN-LDA (*P* = 1.77E-4); for T2D, TVAR achieved the best performance (*P* = 7.01e-20), followed by DeepSEA (*P* = 6.72e-8), Eigen (*P* = 5.37e-6), CADD (*P* = 7.54e-4), DANN (*P* = 1.80e-2), FUN-LDA (*P* = 1.94E-2); for SCZ, the performance of TVAR is the best (*P* = 1.45e-4), followed by Eigen (*P* = 1.97e-4), FUN-LDA (*P* = 2.59e-4), DeepSEA (*P* = 6.73e-4), CADD (*P* = 9.76e-4), DANN (not significant).

Although the sample sizes of the four test scenarios are on the same scale, we found that the performance advantage of TVAR is considerable higher on CAD, BRC, and T2D, but comparable to other algorithms on SCZ. In fact, the low accuracy group of the TVAR scores contains a large number of brain-related tissues, suggesting that the functional prediction of variants on brain-related tissues is more challenging.

**Identification of the tissue-specific functionality of rare variants**

Since the training set of the TVAR were based significance of eQTLs, which are biased towards common variants owing to their statistical power, we are interested in evaluating TVAR's capability in predicting functionality of rare variants, based on the rationale that common and rare regulatory variants share similar regulatory machinery such that they exist similar functional genomics data across tissues and cell types. We constructed four test sets based on the four previous GWAS of CAD, BRC, T2D, and SCZ[31-34]. We constructed the positive sets by including relatively significant rare variants from the four GWAS results with *P* <1e-3 and MAF <0.01. For CAD, BRC, T2D, and SCZ, the numbers of positive sets are 191, 195, 53, 185, respectively. For each test scenario, we randomly selected an equal number of variants from the 1000 Genomes Project as the negative set (MAF <0.01). Similar to the evaluation for common variants, we used one-sided Wilcoxon signed-rank tests to explore whether each algorithm can distinguish rare variants in the positive and the negative sets. From Fig. 3 and Table 2, we observed that only TVAR could significantly distinguish the positive and negative samples of rare variants for all of the four diseases investigated (*P* values are 3.76e-3,

2.60e-6, 2.74e-2, 1.61e-2, for CAD, BRC, T2D, and SCZ, respectively). As a direct contrast, all methods (CADD, Eigen, DeepSEA, DANN) that do not distinguish between tissues, failed to separate the positive and negative samples on all data sets (Fig. 3, Table 2). Even FUN-LDA, which has tissue-specific scores, has challenges to achieve a significant discrimination, with T2D being close to significance with *P* = 7.63e-2.

**Organism-level Functional Predictions with TVAR**

Although TVAR is a tissue-specific functional variant prediction approach, the G-Score scoring algorithm of TVAR can also be used to predict the variant function at the organicism level. To evaluate the performance of the TVAR-GScore algorithm, we constructed four test scenarios: Clinvar (version: 20170130), fine-mapping GWAS hits[26], GTEX eQTLs, and MPRA validated variants[28]. The positive samples of Clinvar include 2,713 clinically validated pathogenic noncoding variants. The positive samples of the fine-mapping GWAS set includes 1,857 fine-mapped GWAS loci across 39 phenotypes. The positive samples of the GTEX eQTLs dataset include highly trusted 1,137 noncoding variants from the GTEx dataset[13]. The positive samples of the MPRA-validated variants dataset include 234 validated noncoding variants reported based on massively parallel reporter assays. The negative samples of the above four test sets were all randomly selected in equal amounts from the variants reported by 1000 Genomes Project (filtered with MAF> 0.05). To ensure the unbiasedness of the test, on the four data sets, we also removed the variants that have appeared in the training set of TVAR. We chose four organicism level methods as comparison algorithms: CADD,

DANN, DeepSEA, and Eigen. On the four data sets, we mainly use ROC curves to evaluate the accuracy of various algorithms (Fig. 4). For comparison, we also give PR curves of each method in the four test scenarios (see Supplementary Note 2).

On the ClinVar dataset, TVAR achieved the best performance (AUC = 0.981), followed by DeepSEA (AUC = 0.975), CADD (AUC = 0.963), Eigen (AUC = 0.958), DANN (AUC = 0.957). The AUCs of all algorithms on the ClinVar data set exceed 0.95, and the gap is not large, indicating that the functional characteristics of the clinically validated deleterious variants are clearly distinguishable so all methods have superior performance. On the Refined GWAS Hits dataset, TVAR obtained the best performance (AUC = 0.733), followed by Eigen (AUC = 0.616), DeepSEA (AUC = 0.615), CADD (AUC = 0.586), DANN (AUC = 0.545). TVAR has an absolute performance improvement of 11.7% ~ 18.8% on this data set. The Refined GWAS Hits dataset contains many functional risk variants of complex diseases, and the clear advantage of TVAR over other method indicates the potential of TVAR in identifying causal variants for complex diseases. On the GTEX eQTLs dataset, TVAR achieved the best performance (AUC = 0.809), followed by DeepSEA (AUC = 0.622), Eigen (AUC = 0.583), CADD (AUC = 0.555), and DANN (AUC = 0.521). Compared to other methods, TVAR has achieved an absolute performance improvement of 18.7% ~ 28.8%. On the MPRA validated variants dataset, TVAR obtained the best performance (AUC = 0.880), followed by DeepSEA (AUC = 0.765), Eigen (AUC = 0.643), CADD (AUC = 0.627), DANN (AUC = 0.598). TVAR has achieved an absolute performance improvement of 11.5% ~ 28.2% over the competing methods. MPRA has validated the variants on this dataset, and the credibility is high. The AUC of

TVAR on this dataset also reaches 0.88, which shows that it can effectively predict MPRA validated functional variants. TVAR's clear margins of improvement over other methods in all scenarios investigated demonstrate its robustness in predicting the functionality of noncoding variants.

## DISCUSSION

We propose a deep neural network-based framework, TVAR, to evaluate the functionality of variants in non-coding regions of the genome. TVAR supports the prediction of tissue-specific functional effects of non-coding variants through multi-label learning. Simultaneously, we combined the output of the TVAR model and used multi-instance learning to perform TVAR's G-score scoring, which supports organicism level functional prediction across tissues. We constructed a large-scale functional variant dataset across 49 tissues (i.e. eQTLs) and the companion functional annotations, which provides a training set suitable for multi-label learning and supervised learning for generating tissue-specific functional prediction. In this set with more than 70,000 variants as the training set, we used 5-fold cross-validation to train and test the TVAR model. We found that in all GTEx tissues, combined with various annotations as features, TVAR achieve an AUC of 0.77 of the functional variants across all the tissues. In brain-related tissues, functional variants that TVAR can explain are generally small (from 69% to 76%). In Esophagus-related tissues, the percentage of variants that can be explained by TVAR is higher (more than 82%). Evaluating TVAR in tissues relevant to 4 complex disease, i.e. CAD, BRC, T2D, and SCZ, we found that the

disease-associated variants have significantly higher TVAR scores than background variants, supporting that the TVAR score can be effective in prioritizing disease-specific functional variants. In the end, the G-Score algorithm based on multi-instance learning achieved excellent performance on four well calibrated functional noncoding variants, clarifying the reliability of the TVAR on the scoring of the functional variants at the organism level.

For functional variants scoring, there are already methods that use deep learning technology: DANN adopts the CADD features and training data and uses the DNN model to replace CADD's SVM classification model; both DeepSEA and ExPecto use deep convolutional networks to encode variants and DNA sequences into feature vectors. However, none of them predict variants at the "multi-tissue" level (i.e., considering the differences and connections among all tissues). To the best of our knowledge, TVAR is the first method that introduces DNN (with multi-label learning) to the multi-tissue level functional variants prediction. We found that in the positive training set of TVAR, more than 30% of the variants have the effect of regulating gene expression in no less than two tissues, and more than 10% of the variants regulate the expression of related genes in no less than four tissues. TVAR uses a large number of fully connected layers in network design. The final output nodes of the model are directly or indirectly related to all the nodes in all the previous layers, which guarantees that TVAR uses non-linear functions to capture the differences and connections among tissues.

Predicting the functional consequences of rare noncoding variants has been a

challenge. With the use of CRISPR-Cas9 and other technologies, some common variants have been studied. However, the functional effects of rare variants and their regulatory mechanisms in corresponding tissues are still elusive. As shown in previous studies[35], randomly selected rare variants also tend to be more likely to be functional than common variants, making it imperative to find real functional rare variants in different tissues to facilitate genetics of common diseases. In this study, in the tissue-specific rare variants test data set, we set the eQTL *p*-value threshold to 1e-3 to guarantee an adequate number of variants in the positive sample set. As a consequence, this positive set is more difficult to separate from the negative set. Among all the methods investigated, only TVAR can distinguish the positive and negative sets significantly, demonstrating TVAR's advantages in learning the subtle effects of rare noncoding variants. This benefits from the principle of TVAR's design, i.e. rare regulatory variants share the same genomic profiles as common functional variants, which enables the transfer of learned relationships of genomic features with common eQTLs to effectively predict the function of rare regulatory variants.

TVAR's model training between tissues is performed at one time, so the number of tissues supported by the TVAR can be expanded with a small computational cost. A specific example is that the TVAR is updated with the GTEx (V7 to V8) and can also support the corresponding new tissues. The earlier exploration of the TVAR was carried out on GTEx V7 (see Supplementary Note 3). We found that on the 48 tissues of the previous version of GTEx, the performance of the TVAR is almost the same as it is now. The network structure used by TVAR is the same as previously. The

performance results illustrate the scalability and stability of the TVAR. In 49 tissues of GTEx (V8), TVAR can achieve about an AUC of 0.77 to classify functional variants. However, we found that there are still many functional variants in brain-related tissues that cannot be explained by the functional annotation of the input features of TVAR. This is not unexpected, however, as brain tissues are highly heterogeneous with diverse cell types with different functions. We also uncovered that the use of multi-instance learning for organicism level functional scoring made significant improvements over existing methods in specific scenarios, which revealed a new direction in the development of organicism level functional scoring algorithms. In the current G-score algorithm, we have adopted the max function. In the future, we will try to use unsupervised algorithms (such as the DPGMM[36] method) to better measure the contribution of different tissues to the organicism level score, to optimize the G-score algorithm.

## METHODS

### Overview of TVAR

The input of TVAR is 1247-dimensional functional annotations of the variants across multiple databases. The output of TVAR is the corresponding functional scores on the 49 tissues in the GTEx project that contain eQTL analysis results (Fig. 5). We preprocess the input features to block the model from overfitting during training. The principal component analysis (PCA) is used to retain the essential information in the input of the TVAR. Next, we obtain the low-dimensional representation of or each variant and apply a deep learning model to classify the preprocessed input. A deep neural network (DNN) is also developed to accomplish the prediction task. The DNN model uses an output layer containing 49 nodes (corresponding to 49 tissues) according to the classification purposes. During the training process, the same variants in the training data are used to represent the functionalities in different tissues, so that they can simultaneously learn whether and in which tissues the variants are functional. TVAR uses multi-label learning rather than multi-class learning methods to predict the functional effects of noncoding variants. TVAR uses multiple fully connected layers and Batch Normalization technique to regularize the mean and variance. The model also uses dropout technique to randomly remove some network nodes and reduce overfitting during network training. The total number of network layers is 23. A well-trained network can learn the patterns of different tissues, and simultaneously output the scores across GTEx 49 tissues (range: 0 ~ 1). To support the scoring in the case of organicism level, we also developed the G-Score method, which connects different

scores of variants in different tissues, and finally outputs an overall functional score.

**The input and features of TVAR**

Many previous studies have shown that functional annotation across multiple databases is essential for variant scoring systems. To assess the functionality of variants as comprehensively as possible, we integrated three variant-related databases of ENCODE, Roadmap, and FANTOM5. We attached four conservative scores of PhastCons[37], PhyloP[38], GERP++[39] and SiPhy[40] , and Transcription factor (TF) binding sites, super-enhancer regions, transcription start site (TSS), CpG islands and other information to provide functional evidence to support variants. Specifically, from the ENCODE database, we selected the labels related to TF binding sites (132 values), Histone modifications labels (18 values), DNase clusters labels (1 value); from the Roadmap database, we selected labels such as DNase.macs2 (53 values), H2A.Z (23 values), H2AK5ac (7 values), H2AK9ac (1 values), H2BK120ac (7 values), H2BK12ac (6 values), H2BK15ac (5 values), H2BK20ac (3 values), H2BK5ac ( 7 values), H3K14ac (6 values), H3K18ac (7 values), H3K23ac (7 values), H3K23me2 (2 values), H3K27ac (98 values), H3K27me3 (127 values), H3K36me3 (127 values), H3K4ac (7 values) ), H3K4me1 (127 values), H3K4me2 (24 values), H3K4me3 (126 values), H3K56ac (3 values), H3K79me1 (7 values), H3K79me2 (21 values), H3K9ac (62 values), H3K9me1 (4 values), H3K9me3 (127 values), H3T11ph (1 values), H4K12ac (1 values), H4K20me1 (19 values), H4K5ac (3 values), H4K8ac (7 values), H4K91ac (6 values). From the FANTOM 5 database, we select CAGE peaks (1 value), permissive enhancers (1 value),

494 robust enhancers (1 value). From the 4DGenome[41] database, we choose chromatin
495 interactions detected maker (1 value). From some other related studios, we collect
496 super-enhancers[42] (1 value), CpG islands maker (1 value), the distance of the variant
497 to the nearest TSS (1 value). In addition, we also integrate PhastCons score (3 values),
498 PhyloP score (3 values), GERP ++ score (2 values), SiPhy score (1 values). Since variants
499 do not necessarily implicate their nearest genes[43,44], for each of them, we calculate
500 the average gene expression (49 values) of the two closest genes in 49 tissues.

501 For each variant, there are a total of 1247 input features. Over 1,000 functional
502 annotations enable the TVAR framework to comprehensively learn the functional
503 similarities and differences between variants in different tissues. The total number of
504 tissue-specific features in the TVAR framework is 1098, accounting for 88% of all
505 features. TVAR integrates these tissue-specific features and measures genome-wide
506 functional scores of variants across different tissues with deep learning technology. Of
507 all the inputs, 1188 values are bool-type data. For each variant, we use PCA to reduce
508 these 1188 values to 96 values and merge the other 59 values to obtain a 155-
509 dimensional vector in the training and test sets. Next, we use DNN to model the
510 various input annotations. To construct supervised learning labels on the training set,
511 for all variants in the training set, we determine the training labels of the variants in
512 the dataset based on the q-values (q-value < 0.01) of eQTLs in the GTEx tissues and the
513 LINSIGHT score (top 30%). A variant must meet these two thresholds concurrently to
514 be selected into the training set. Finally, the labels of the training set data are used for
515 the supervised multi-label learning DNN model.

**The DNN model of TVAR**

TVAR is designed to be a multi-label learning approach that measures the functional effects of variants across tissues to learn the knowledge of the correlation among tissues. As shown in Fig. 5, the input of the network is vector **x**, corresponding to all the functional annotations concatenated, for each variation. This information is passed from the input layer to the output layer. The output is a 49-dimensional vector, corresponding to 49 tissues in GTEx. Firstly, we use multiple fully connected layers to transform the input **x**:

$$z_j = \sum w_{ij} x_i + b_j$$
$$y_j = f\left(z_j\right) \quad (1)$$

Where i, j, and k represent the indexes of the nodes of each layer in the network (including the input layer, hidden layer, and output layer). Let $\mathbf{x}_i$ be the input of node i, and $\mathbf{y}_j$ be the output of the corresponding hidden node j in the next layer, and $\mathbf{z}_j$ be the value after the transformation. The final output layer is a 49-dimensional vector **y**, and f() denotes a non-linear activation function. We use the ReLU function between the input layer and the hidden layer (and also between the hidden layers):

$$\mathrm{ReLU} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2)$$

In addition, to prevent a large number of fully connected layers from overfitting the model, we also added batch normalization layers and dropout layers between the fully connected layers. The batch normalization layer performs the following operations on each mini-batch B = [x, …, xm]:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \mu_B\right)^2$$
$$x_i^{norm} = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (3)$$
$$y_i = \gamma x_i^{norm} + \beta$$

Where $\mu_B$ denotes the mean of the mini-batch, $\sigma_B$ denotes the variance of the mini-batch, $\mathbf{x}_i^{norm}$ denotes the normalized input, γ and β denote the scale and shift of $\mathbf{x}_i^{norm}$ to get the output $\mathbf{y}_i$. Similarly, the dropout layer is denoted as:

$$r_i \sim Bernoulli\left(p\right)$$
$$y_i = r_i x_i \quad (4)$$

Where the Bernoulli function is used to generate a vector **r** that is then multiplied to the input **x.** The resulting $y_i$ is set to 0 with *p* as the probability. Unlike the multi-class classification task with neural networks, the loss function of the DNN model of TVAR is designed to support the Sigmoid Cross-Entropy loss of multi-label learning:

$$\mathrm{L} = -\frac{\sum_{\mathrm{i}=1}^{m}\sum_{j=1}^{n}\left[\begin{array}{l} l_{i,j}\log f\left(x_{i,j}\right) \\ +\left(1-l_{i,j}\right)\log\left(1-f\left(x_{i,j}\right)\right)\end{array}\right]}{mn} \quad (5)$$

Where $x_{i,j}$ denotes the *i*th output node (m nodes in total) of the *j*th variant (a total of n variants) in the data set. Let *l* be the training label of (sample, issue), and L be the total loss, *f* () is the sigmoid function:

$$f\left(x_{i,j}\right) = \frac{1}{1+e^{-x_{i,j}}} \quad (6)$$

The specific network structure of the DNN model is shown in Table S1. The network parameters, learning rate, and optimization methods are fitted during the training process. After the training of the model, for each variant **x**, a score vector **y**

corresponding to the 49 GTEx tissues is generated.

**The G-score algorithm of TVAR**

Although the design of the TVAR is to score the variant-tissue pairs. However, with the development of the G-score algorithm, we can make the TVAR support variant scoring at the organism level. For a score that does not distinguish between issues, only one score s is needed for each variation. The G-score algorithm is designed as a multi-instance learning approach: the score **y** on 49 distinct tissues for each variant **x** is regarded as a bag. For bag **y**, we use a function f() to find its general score:

$$s = f(\mathbf{y}) \tag{7}$$

In this study, we simply consider the maximum score among tissues as the G-score for each variant, i.e. f () takes the max function.

**Availability of Code and Functional Scores of TVAR**

The TVAR source code and its scores on the ClinVar catalog, fine mapped GWAS Loci, high-confidence eQTLs from GTEx dataset, and MPRA validated functional variants are available at https://github.com/haiyang1986/TVAR.

**Acknowledgements**

This work was partially supported by NIH grant U01HG009086 to RC, QWa, JY, QWe, XZ and BL, as well as to HY during the period when HY was a Postdoctoral Fellow at Vanderbilt University.

## Author contributions

BL and HY designed the study. HY developed the TVAR and implemented it. RC, QWa, JY, QWe and XZ collated functional annotations and contributed to the interpretation of the results. HY and BL wrote the manuscript. All authors read and approved the manuscript.

## Competing financial interests

None of the authors declare competing financial interests.

## Reference


1 Human Microbiome Project, C. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214, doi:10.1038/nature11234 (2012).

2 Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110, doi:10.1093/hmg/ddv259 (2015).

3 Gloss, B. S. & Dinger, M. E. Realizing the significance of noncoding functionality in clinical genomics. *Exp Mol Med* **50**, 97, doi:10.1038/s12276-018-0087-0 (2018).

4 Gulko, B. & Siepel, A. An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences. *Nature genetics* **51**, 335-342, doi:10.1038/s41588-018-0300-z (2019).

5 Skipper, M., Dhand, R. & Campbell, P. Presenting ENCODE. *Nature* **489**, 45, doi:10.1038/489045a (2012).

6 Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350-354, doi:10.1038/nature14217 (2015).

7 Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* **16**, 22, doi:10.1186/s13059-014-0560-6 (2015).

8 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

9 Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-296, doi:10.1038/nmeth.2832 (2014).

10 Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761-763, doi:10.1093/bioinformatics/btu703 (2015).

11 Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics*, doi:10.1038/ng.3810 (2017).

12 Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature genetics* **48**, 214-220, doi:10.1038/ng.3477 (2016).

13 Yang, H. *et al.* De novo pattern discovery enables robust assessment of functional consequences of non-coding variants. *Bioinformatics* **35**, 1453-1460, doi:10.1093/bioinformatics/bty826 (2019).

14 Bodea, C. A. *et al.* PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. *Genome biology* **19**, 173, doi:10.1186/s13059-018-1546-6 (2018).

15 Janssens, A. C. *et al.* The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet Med* **9**, 528-535, doi:10.1097/gim.0b013e31812eece0 (2007).

16 Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience* **19**, 48-+, doi:10.1038/nn.4182 (2016).

17 Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 17921-17926, doi:10.1073/pnas.1317023110 (2013).

18 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

19 Consortium, G. T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213, doi:10.1038/nature24277 (2017).
20 Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics* **47**, 955-961, doi:10.1038/ng.3331 (2015).
21 He, Z. H., Liu, L. X., Wang, K. & Ionita-Laza, I. A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nature Communications* **9**, doi:10.1038/s41467-018-07349-w (2018).
22 Backenroth, D. *et al.* FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am J Hum Genet* **102**, 920-942, doi:10.1016/j.ajhg.2018.03.026 (2018).
23 Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics* **50**, 1171-1179, doi:10.1038/s41588-018-0160-6 (2018).
24 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016).
25 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980-D985, doi:10.1093/nar/gkt1113 (2014).
26 Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343, doi:10.1038/nature13835 (2015).
27 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).
28 Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529, doi:10.1016/j.cell.2016.04.027 (2016).
29 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291-295, doi:10.1038/ng.3211 (2015).
30 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
31 Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature genetics* **47**, 1121-1130, doi:10.1038/ng.3396 (2015).
32 Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun* **9**, 2941, doi:10.1038/s41467-018-04951-w (2018).
33 Pardinas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* **50**, 381-389, doi:10.1038/s41588-018-0059-2 (2018).
34 Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics* **47**, 373-380, doi:10.1038/ng.3242 (2015).
35 Yang, H. *et al.* De Novo pattern discovery enables robust assessment of functional consequences of noncoding variants. *Bioinformatics*, doi:10.1093/bioinformatics/bty826 (2018).
36 Yang, H., Wei, Q., Zhong, X., Yang, H. & Li, B. Cancer driver gene discovery through an

integrative genomics approach in a non-parametric Bayesian framework. *Bioinformatics* **33**, 483-490, doi:10.1093/bioinformatics/btw662 (2017).

37 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

38 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* **15**, 901-913, doi:10.1101/gr.3577405 (2005).

39 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

40 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62, doi:10.1093/bioinformatics/btp190 (2009).

41 Teng, L., He, B., Wang, J. & Tan, K. 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* **31**, 2560-2564, doi:10.1093/bioinformatics/btv158 (2015).

42 Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947, doi:10.1016/j.cell.2013.09.053 (2013).

43 Wang, Q. *et al.* A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat Neurosci* **22**, 691-699, doi:10.1038/s41593-019-0382-7 (2019).

44 Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-375, doi:10.1038/nature13138 (2014).

**FIGURES**

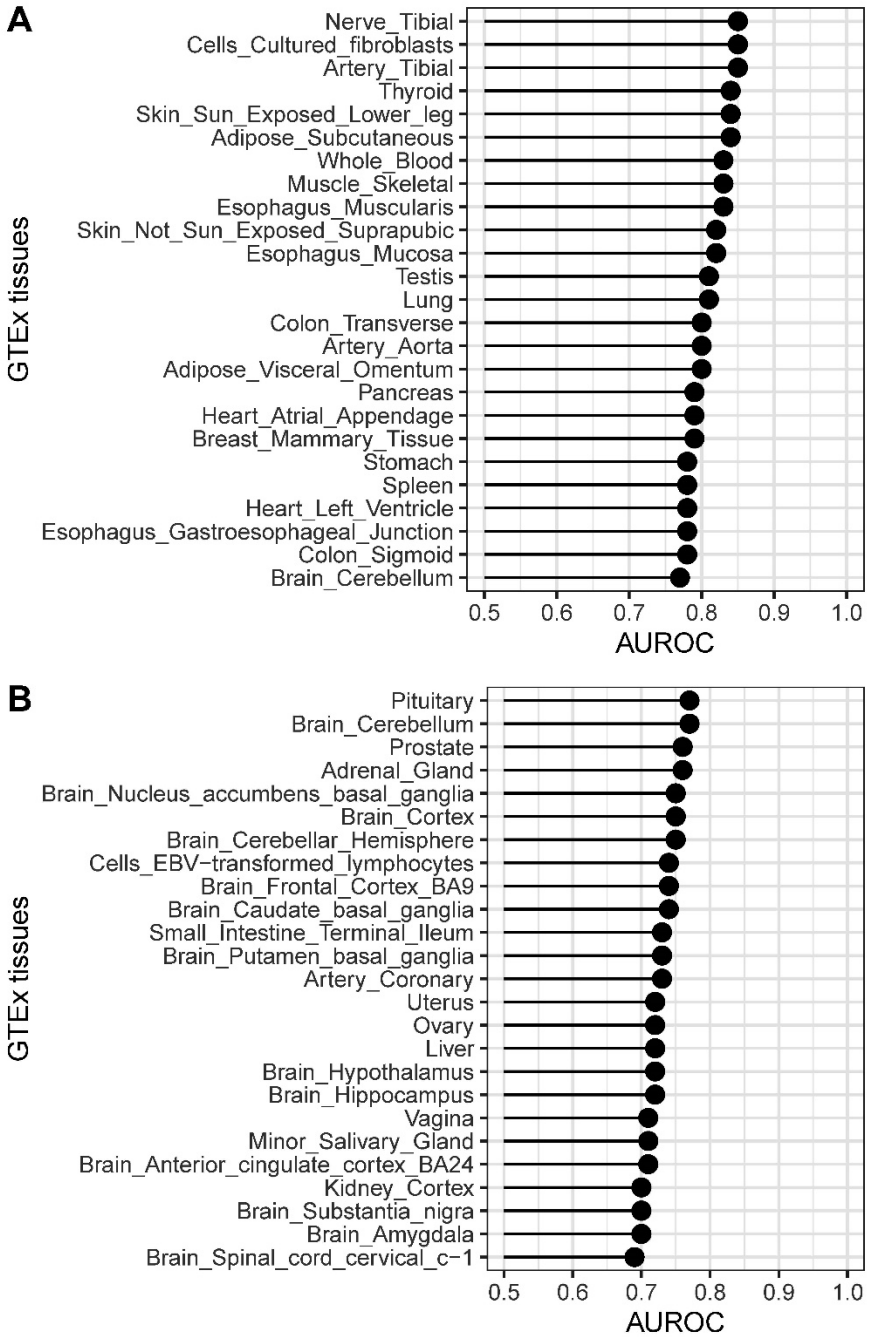



Fig. 1 Five-Fold cross-validation performance of TVAR on the GTEx 49 tissues: A). High accuracy group. B) Low accuracy group.

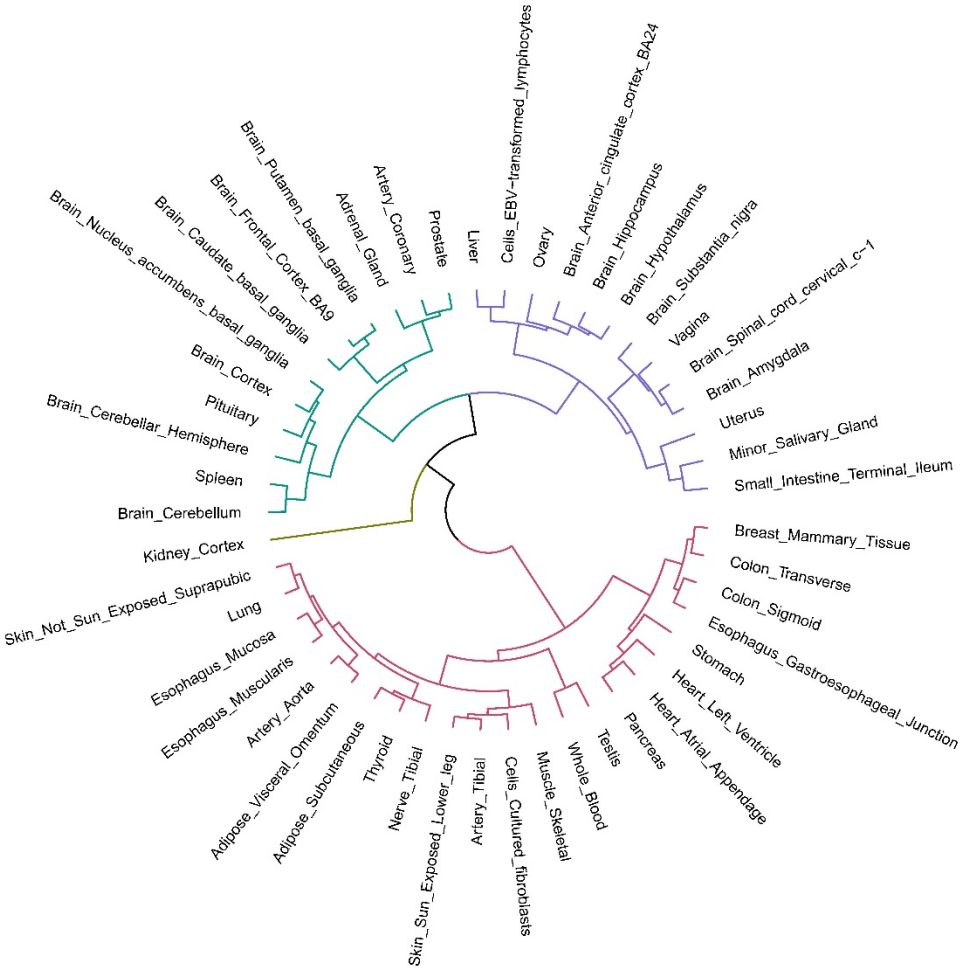



Fig. 2: The hierarchical clustering results of TVAR's functional scores across the GTEx 49 tissues (from the GTEx dataset V8).

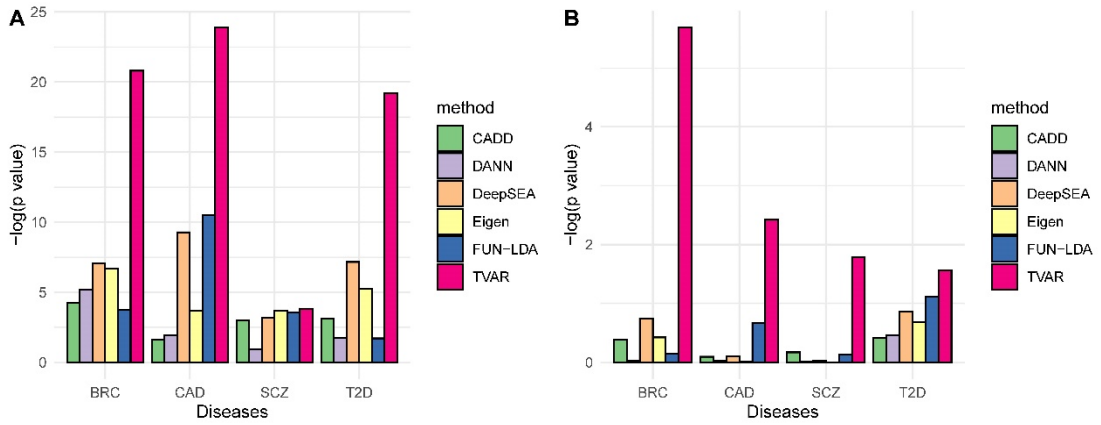



Fig. 3 Performance comparison of TVAR, CADD, DANN, DeepSEA, Eigen, FUN-LDA on datasets of four complex diseases: coronary artery disease (CAD), breast cancer (BRC), Type 2 diabetes (T2D), and Schizophrenia (SCZ): (A) GWAS hit data set; (B)rare variant data set. The prediction power was evaluated based on the -log10(p-value).

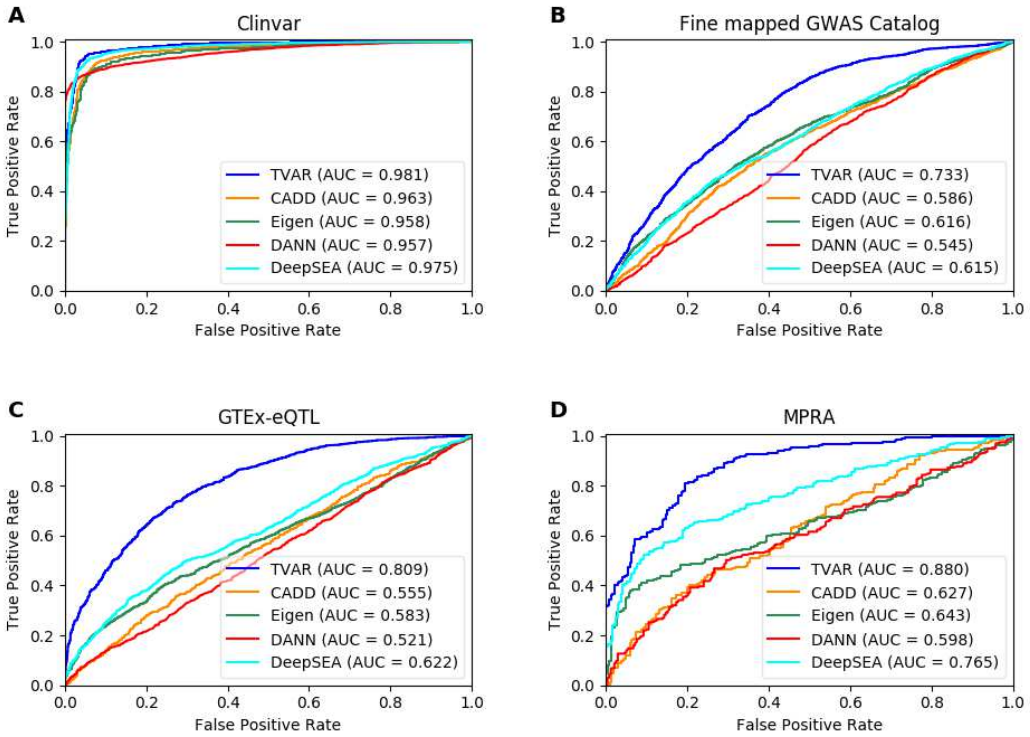
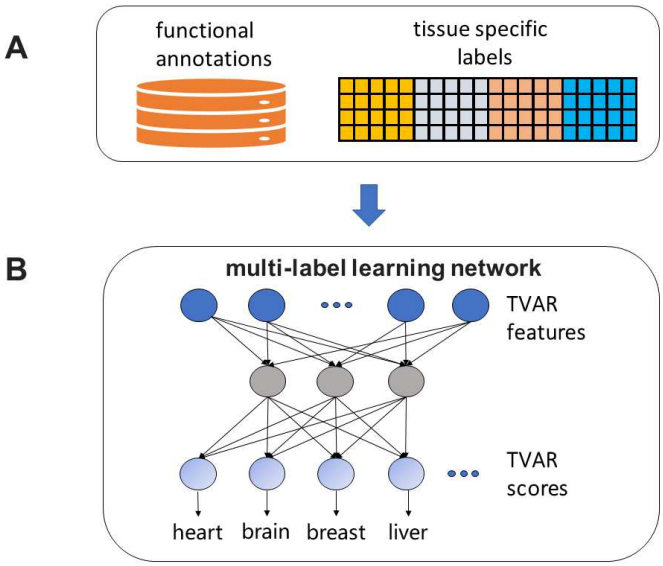



Fig. 4 Performance comparison of TVAR, CADD, DANN, DeepSEA, and Eigen on the extended dataset: A) clinically significant variants in Clinvar database, B) fine-mapped GWAS variants), C) GTEx-eQTLs, and D) MPRA validated variants. The identification accuracy is counted with the area under the ROC curves.




C

| Tissue | Variant | TVAR score |
|---|---|---|
| Heart | rs2843152,<br>rs3609619 | 0.072<br>0.008 |
| Brain | rs3001723 | 0.041 |
| ... | ... | ... |

Fig. 5 Overview of the TVAR framework. (A) Take the functional annotations as the input of the deep learning network. (B) Multi-label learning with DNN. (C) Output the functional scores across multiple tissues.

Table 1 The p-values (Wilcoxon signed-rank test, one-sided with 'greater') for the analysis of common variants on four diseases (coronary artery disease, breast cancer, Type 2 diabetes, and Schizophrenia):

| Disease | N | Score | p-value |
|---|---|---|---|
| coronary artery disease (cardiovascular) | 342 | TVAR | **1.21E-24** |
| | | CADD | **2.30E-2** |
| | | Eigen | **1.93E-4** |
| | | DANN | **1.14E-2** |
| | | DeepSEA | **5.29E-10** |
| | | FUN-LDA | **3.25E-11** |
| breast cancer (cancer) | 354 | TVAR | **1.60E-21** |
| | | CADD | **5.52E-5** |
| | | Eigen | **2.02E-7** |
| | | DANN | **6.20E-6** |
| | | DeepSEA | **8.20E-08** |
| | | FUN-LDA | **1.77E-4** |
| Type 2 diabetes (metabolic disease) | 210 | TVAR | **7.01E-20** |
| | | CADD | **7.54E-4** |
| | | Eigen | **5.37E-6** |
| | | DANN | **1.80E-2** |
| | | DeepSEA | **6.72E-8** |
| | | FUN-LDA | **1.94E-2** |
| Schizophrenia (mental) | 286 | TVAR | **1.45E-4** |
| | | CADD | **9.76E-4** |
| | | Eigen | **1.97E-4** |
| | | DANN | 1.13E-1 |
| | | DeepSEA | **6.73E-4** |
| | | FUN-LDA | **2.59E-4** |

Table 2. The p-values (Wilcoxon signed-rank test, one-sided with 'greater') for the analysis of rare variants on four diseases (coronary artery disease, breast cancer, Type 2 diabetes, and Schizophrenia):

| Disease | N | Score | p-value |
|---|---|---|---|
| coronary artery disease (cardiovascular) | 191 | TVAR | **3.76E-3** |
| | | CADD | 8.04E-1 |
| | | Eigen | 9.71E-1 |
| | | DANN | 9.26E-1 |
| | | DeepSEA | 7.92E-1 |
| | | FUN-LDA | 2.16E-1 |
| breast cancer (cancer) | 195 | TVAR | **2.06E-6** |
| | | CADD | 4.14E-1 |
| | | Eigen | 3.78E-1 |
| | | DANN | 9.39E-1 |
| | | DeepSEA | 1.79E-1 |
| | | FUN-LDA | 7.19E-1 |
| Type 2 diabetes (metabolic disease) | 53 | TVAR | **2.74E-2** |
| | | CADD | 3.86E-1 |
| | | Eigen | 2.05E-1 |
| | | DANN | 3.44E-1 |
| | | DeepSEA | 1.36E-1 |
| | | FUN-LDA | 7.63E-2 |
| Schizophrenia (mental) | 189 | TVAR | **1.61E-2** |
| | | CADD | 6.74E-1 |
| | | Eigen | 9.93E-1 |
| | | DANN | 9.69E-1 |
| | | DeepSEA | 9.33E-1 |
| | | FUN-LDA | 7.42E-1 |

# Figures

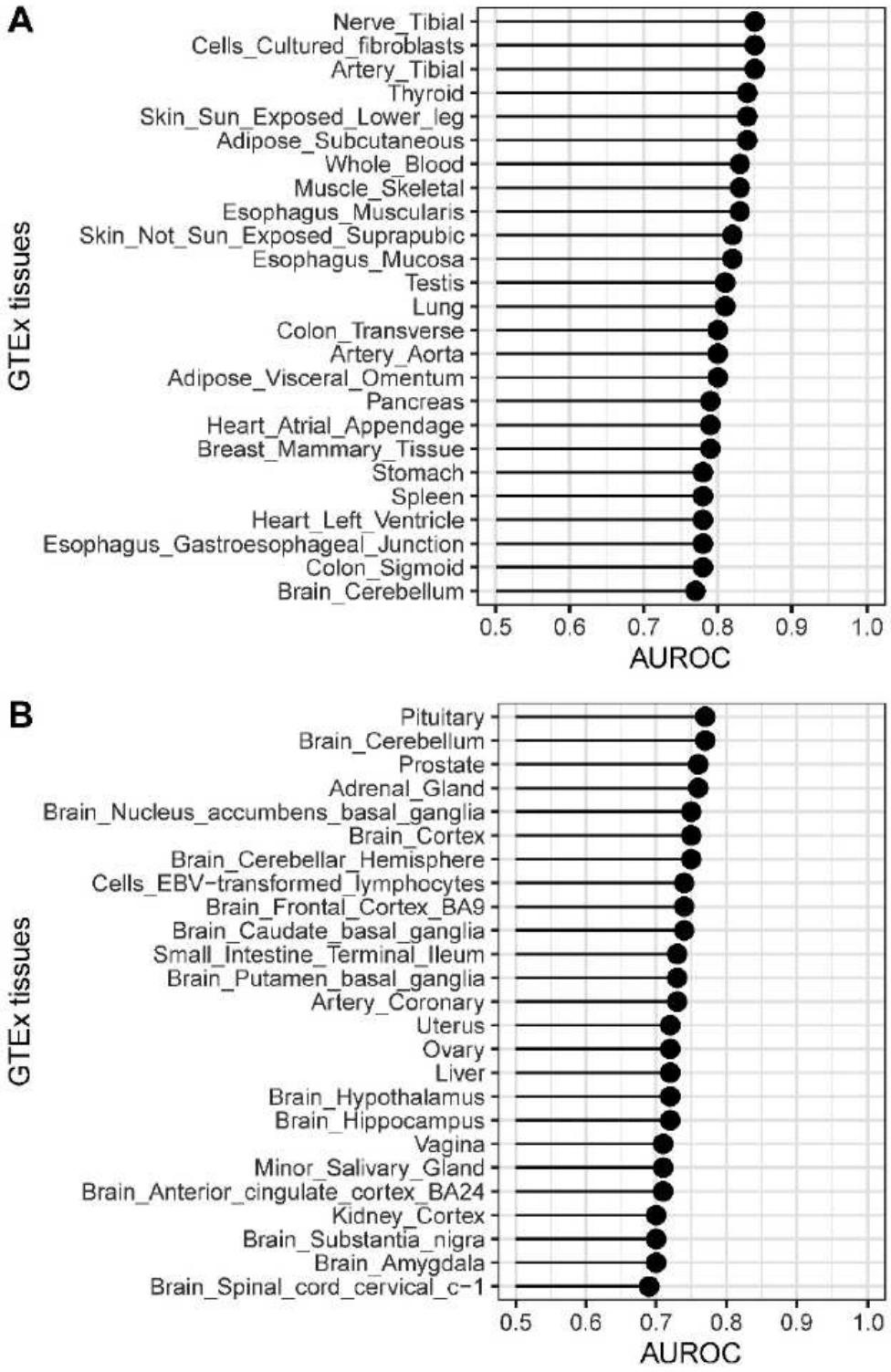
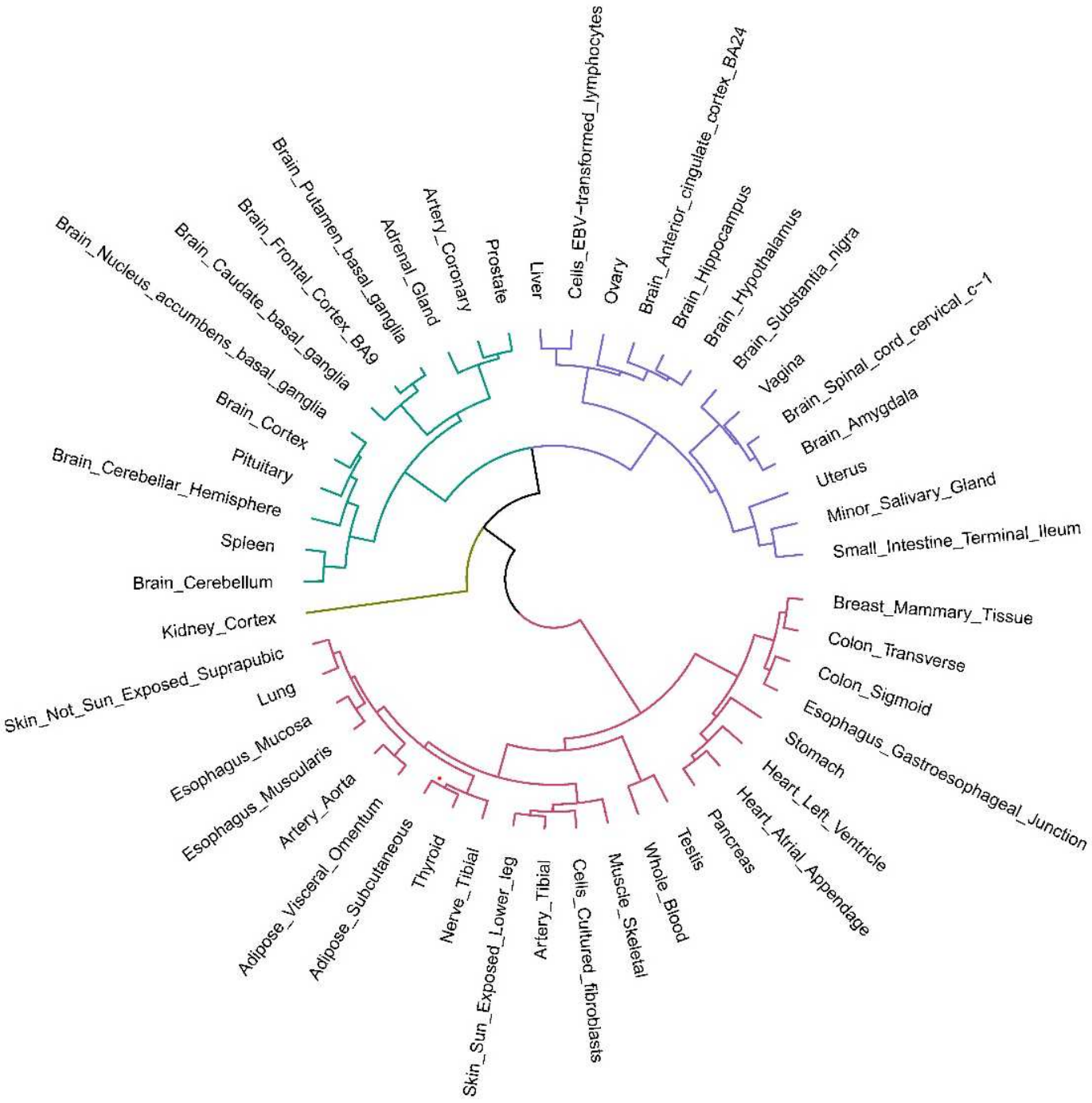


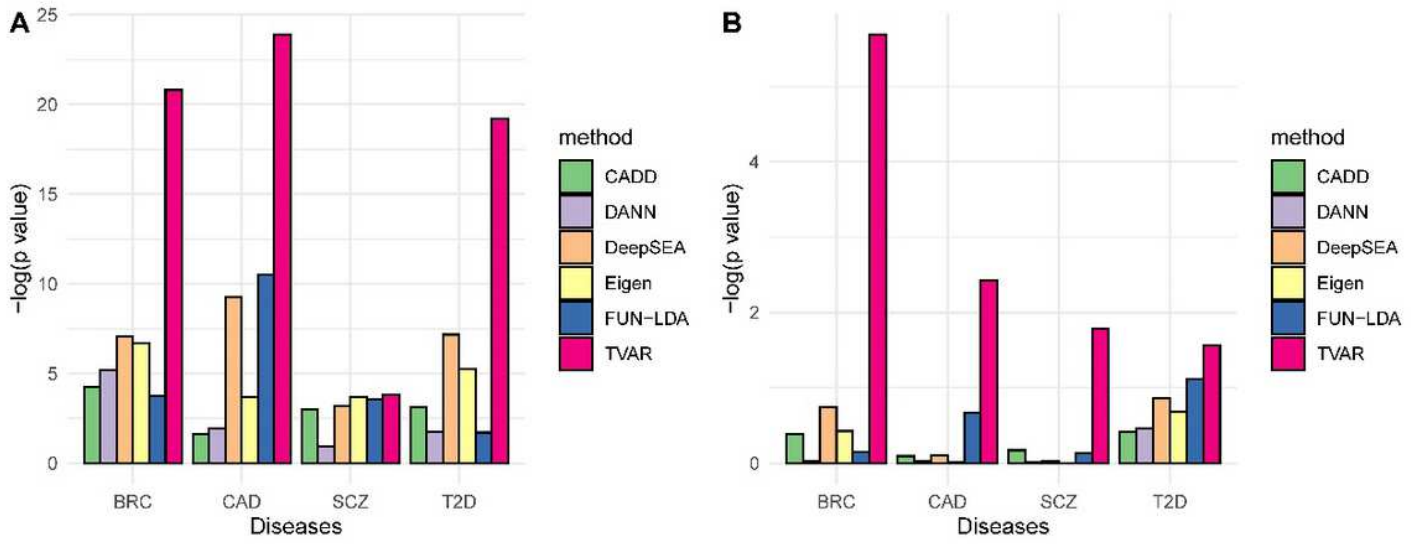

**Figure 1**

Five-Fold cross-validation performance of TVAR on the GTEx 49 tissues: A). High accuracy group. B) Low accuracy group.




## Figure 2

The hierarchical clustering results of TVAR's functional scores across the GTEx 49 tissues (from the GTEx dataset V8).




## Figure 3

Performance comparison of TVAR, CADD, DANN, DeepSEA, Eigen, FUN-LDA on datasets of four complex diseases: coronary artery disease (CAD), breast cancer (BRC), Type 2 diabetes (T2D), and Schizophrenia (SCZ): (A) GWAS hit data set; (B)rare variant data set. The prediction power was evaluated based on the -log10(p-value).

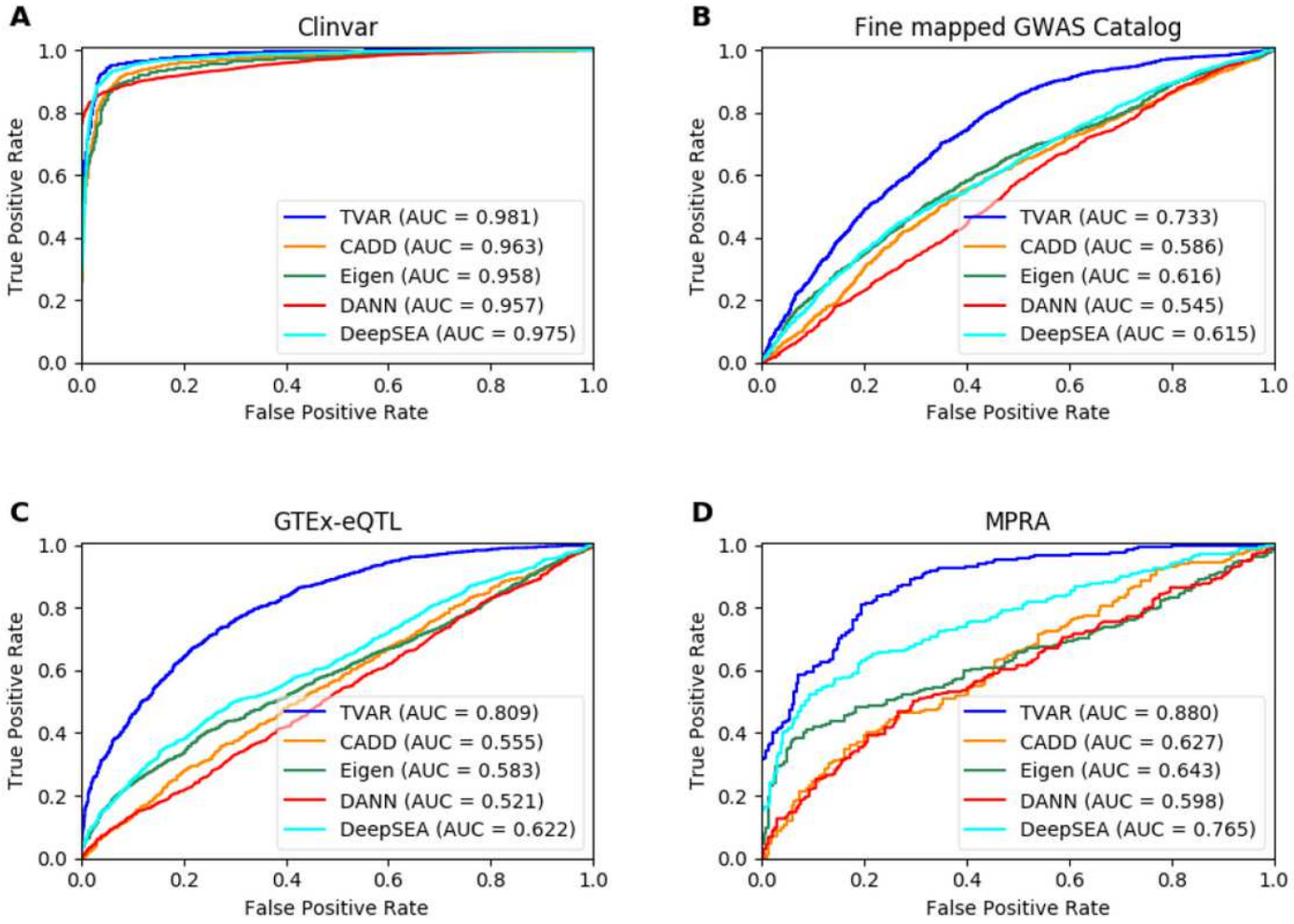
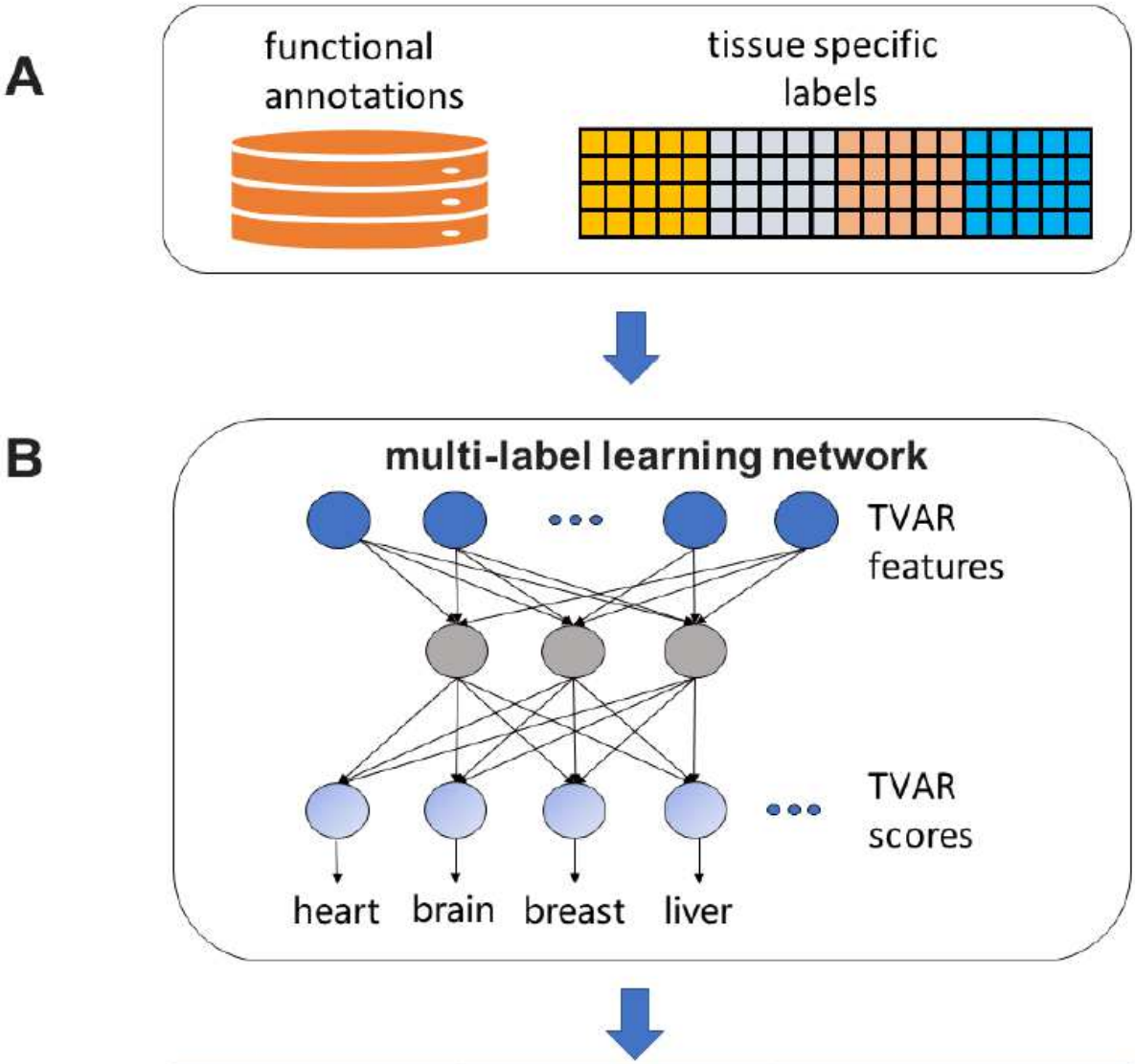



**Figure 4**

Performance comparison of TVAR, CADD, DANN, DeepSEA, and Eigen on the extended dataset: A) clinically significant variants in Clinvar database, B) fine-mapped GWAS variants), C) GTEx-eQTLs, and D) MPRA validated variants. The identification accuracy is counted with the area under the ROC curves.




C

| Tissue | Variant | TVAR score |
|---|---|---|
| Heart | rs2843152,<br>rs3609619 | 0.072<br>0.008 |
| Brain | rs3001723 | 0.041 |
| ... | ... | ... |

**Figure 5**

Overview of the TVAR framework. (A) Take the functional annotations as the input of the deep learning network. (B) Multi-label learning with DNN. (C) Output the functional scores across multiple tissues.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 02TVARsupplementarymaterials20201121.pdf