Genomics Proteomics Bioinformatics 21 (2023) 662–673

Genomics Proteomics Bioinformatics

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

METHOD

# Computational Assessment of the Expression-modulating Potential for Non-coding Variants

**Fang-Yuan Shi [1], Yu Wang [1], Dong Huang [2], Yu Liang [3], Nan Liang [1], Xiao-Wei Chen [2,4], Ge Gao [1,*]**

[1] *State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Biomedical Pioneering Innovative Center (BIOPIC) & Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), Peking University, Beijing 100871, China*
[2] *State Key Laboratory of Membrane Biology, Institute of Molecular Medicine, Peking University, Beijing 100871, China*
[3] *Human Aging Research Institute, School of Life Science, Nanchang University, Nanchang 330031, China*
[4] *Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*

Received 27 January 2021; revised 13 October 2021; accepted 1 November 2021
Available online 7 December 2021

Handled by Yi Xing

**KEYWORDS**

Non-coding variant;
Expression-modulating variant;
Gene regulation;
Algorithm;
Web server

**Abstract** Large-scale genome-wide association studies (GWAS) and expression quantitative trait locus (eQTL) studies have identified multiple **non-coding variants** associated with genetic diseases by affecting gene expression. However, pinpointing causal variants effectively and efficiently remains a serious challenge. Here, we developed CARMEN, a novel **algorithm** to identify functional non-coding **expression-modulating variants**. Multiple evaluations demonstrated CARMEN's superior performance over state-of-the-art tools. Applying CARMEN to GWAS and eQTL datasets further pinpointed several causal variants other than the reported lead single-nucleotide polymorphisms (SNPs). CARMEN scales well with the massive datasets, and is available online as a **web server** at http://carmen.gao-lab.org.

## Introduction

Approximately 98% of the human genome does not encode proteins [1], and more than 90% of disease-associated variants identified by association studies are non-coding [2–4]. However, their biological functions and mechanisms remain largely elusive [5–8].

* Corresponding author.
E-mail: gaog@mail.cbi.pku.edu.cn (Gao G).

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

https://doi.org/10.1016/j.gpb.2021.10.003

Several algorithms have been developed to prioritize functional non-coding variants based on existing annotations [9–18]. Recently, convolutional neural networks (CNNs) [19] have been employed to characterize the regulatory activity of genomic sequences and call variants which change chromatin profiles [*e.g.*, transcription factor (TF) binding and histone modifications] [20–22], and to estimate variant impact on gene expression *ab initio* [23]. Meanwhile, massively parallel reporter assay (MPRA) offers direct assessment of transcriptional activities for millions of *cis*-elements by transfecting cells with plasmids harboring synthetic elements and reporter genes [24,25], enabling systematic screening for potential regulatory variants [25,26]. Trained on MPRA data, EnsembleExpr successfully identifies variants which change reporter expression significantly, and pinpoints causal variants in several expression quantitative trait locus (eQTL) datasets [27].

Here, we propose CARMEN, a novel algorithm for predicting the effects of non-coding variants on both gene expression and disease risk (**Figure 1**A). Compared with state-of-the-art tools, CARMEN shows superior performance on both high-throughput datasets and low-throughput case studies. In particular, CARMEN's high sensitivity enables effective identification of multiple causal expression-modulating variants that other tools missed. Of interest, CARMEN successfully identified multiple causal variants other than the reported lead single-nucleotide polymorphisms (SNPs) in various datasets of genome-wide association studies (GWAS) and eQTL studies. CARMEN is available as a web server with free access for academic usage at http://carmen.gao-lab.org.

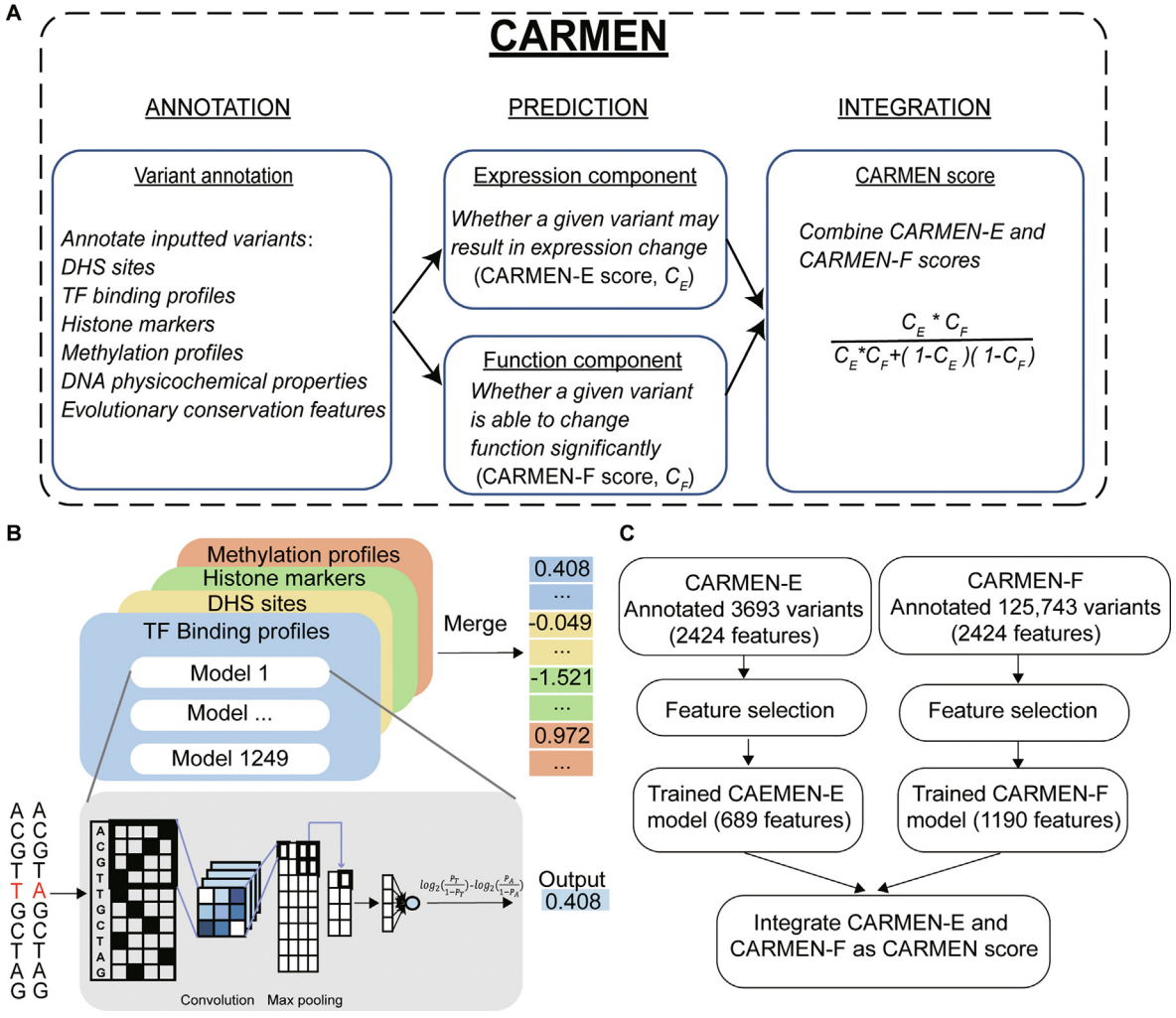


**Figure 1 Schematic overview of CARMEN**

**A.** Overview of the CARMEN workflow. **B.** Pipeline for predicting variant effects on chromatin profiles with sequence-based CNNs. The input data are one-hot encoding data. The first layer is a convolution layer for detecting sequence features, and the max-pooling layer reduces the dimensions of the data. Then, after two fully connected layers, the probability between 0 and 1 is generated, which represents the predicted effect in terms of the $\log_2$ fold change between reference and alternative sequences. **C.** The overall process of CARMEN model training and feature selection. CNN, convolutional neural network; DHS, DNase I-hypersensitive; TF, transcription factor.

## Method

### Annotating variants with multi-modal features

*Predicting variant impacts on chromatin profiles with sequence-based neural networks*

Inspired by pioneering works [22,28], we employed convolutional networks to assess inputted variants' effects on chromatin profiles, including TF binding, histone modification, DNase I-hypersensitive (DHS) sites, and DNA methylation sites. Different from the "one holistic network design" used previously [22], we employed a "one-feature-one-network" design, training independent network for each chromatin profile feature separately for better flexibility and scalability (Figure 1B). The basic unit of the network is a stack of Convolution, Pooling, and Dropout: one 1D convolutional layer with ReLU activation function (kernel number = 128 or 256, kernel length = 4, 10, 12, or 20), followed by one 1D Pooling layer (max pooling, size = 10 or 20) and one Dropout layer (drop rate = 0.2). In case of potential overfitting, a fully connected layer (ReLU activation function) and corresponding Dropout layer (drop rate = 0.5) were introduced afterward, rightly before the final output producing layer (fully connected with sigmoid activation function).

To train these networks, we downloaded raw data for 1249 TF binding profiles, 766 histone markers, 280 DHS sites, and 108 DNA methylation profiles from ENCODE (Table S1), and took data cleaning according to the official guideline (https://www.encodeproject.org/data-standards/): for TF binding data, only those involved in conservative irreproducible discovery rate (IDR) peaks and optimal IDR peaks were kept; for histone marker profiles, replicated peaks were kept; for DHS sites, pseudoreplicated IDR peaks were used. Multiple data files produced by different experiments or labs for the same feature were merged based on genomic coordinates (when two peaks are found to be overlapped, the one with higher peak score will be kept, see Figure S1 for more details). Meanwhile, negative cases were generated as follow: for each TF binding, histone marker, and DHS dataset, we firstly split reference genome with positive data excluded into 200-bp bins, and then randomly sampled these bins to match the number of positive cases; for each methylation dataset, we took sites with methylation rate lower than 50% as negative cases [29]. During training, we employed the binary cross entropy (BCE) as the loss function:

$$BCE(t,p) = -(t \times ln(sigmoid(p)) + (1-t) \times ln(1-sigmoid(p))) \quad (1)$$

where $t$ is the true label and $p$ is the network output. Grid searching was performed to optimize all hyperparameters.

Each network was trained and evaluated with five-fold cross-validation independently. Briefly, in each iteration, we first randomly sampled 15% of data as the independent testing set that was not involved in follow-up training, and then split the rest as the nonoverlapping training set (70%) and the validation set (15%). The most accurate network over five folds was chosen, and the corresponding accuracy was reported.

In efforts to verify that these networks have "learned" genuine information, we inspected the trained networks for known functional motifs. After checking database JASPAR (2020 CORE_vertebrates nonredundant) [30] and TRANSFAC (2019.3 Professional version) [31], we pinpointed 1395 curated TF binding profiles for 357 TFs. We further extracted representative sequence motifs from kernels of corresponding 803 trained networks via the procedure described by Kelley and his colleagues [28]. A direct comparison via Tomtom in the MEME suite [32] identified statistically significant matches ($q < 0.05$) for most (72.10%, or 579 of 803 in total) CARMEN networks with 844 curated TF binding profiles for 220 TFs, strongly suggesting that corresponding networks learn *bona fide* motifs instead of trivial features (see Table S2 for a detailed list and Figure S2 for a few examples).

For each variant, 100-bp flanking sequence upstream and 99-bp flanking sequence downstream of the particular allele will be extracted, resulting in a 200-bp reference sequence and a 200-bp alternative sequence. The reference sequence and alternative sequence will be fed into network then, and the variant effect is estimated as the $\log_2$ fold change of the network output for reference ($P_r$) and alternative sequence ($P_a$):

$$log_2\left(\frac{P_r}{1-P_r}\right) - log_2\left(\frac{P_a}{1-P_a}\right) \quad (2)$$

*Predicting variant impacts on DNA physicochemical properties and extracting evolutionary conservation features at the variant locus*

Following the published protocol [33,34] (see https://github.com/gao-lab/CARMEN/blob/master/06_12features_annotation.py and https://github.com/gao-lab/CARMEN/blob/master/07_OHfeature_annotation.py for more details), we further incorporated the change in 13 DNA physicochemical properties which are shown to be important for regulatory functionality. Meanwhile, we extracted the evolutionary conservation score at the variant locus from the PhastCons and PhyloP tracks for primate, mammal, and vertebrate clades at UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/). LiftOver was used to convert the genome coordinates to GRCh38. In total, we annotated each inputted variant with 2424 features in the CARMEN Annotation Component (Figure S3).

### Training and evaluating CARMEN prediction components

Two distinct components were trained to identify functional non-coding expression-modulating variants: CARMEN-E, for assessing variant impacts on gene expression, and CARMEN-F, for assessing variants' disease risks.

*Dataset curation*

We compiled a MPRA-based dataset for CARMEN-E training by manually curated four recently published peer-reviewed papers, consist of 1082 positive and 2611 negative cases, respectively (Table S3). For training CARMEN-F component, we extracted 1405 disease-causing regulatory variants from HGMD professional version 2018.4 as positive cases, and 124,338 negative cases from 1000 Genomes by following the previous protocol [20] (see Table S4 for more details).

*Data-driven feature selection*

To reduce the risk of overfitting and high time costs of model training, we adopted a data-driven feature selection approach by pretraining multilayer perceptron (MLP) networks with 2424 annotated features as input (Figure S4). The MLP network was trained with optimizer 'Adadelta' (batch size is 100, 100 epochs with early-stopping patience = 30). To extract important features, we used Deep Learning Important FeaTures (DeepLIFT) to estimate features' contribution scores by comparing the difference in the activation value of each neuron to its 'reference activation'. Here, we chose "all zeros" as the reference. After that, we calculated the distribution of the contribution scores of the input data and selected the features with contribution scores higher than the cutoff (see https://github.com/gao-lab/CARMEN-Figures_and_Tables/tree/main/Feature-selection for more details). Finally, we selected 689 and 1190 features for CARMEN-E and CARMEN-F, respectively.

*Training and evaluating CARMEN-E and CARMEN-F*

Given the limited resource, we implemented CARMEN-E as AdaBoost decision tree, but CARMEN-F with less time-consuming Random Forest due to its larger size (1190 features over 125,743 variants *vs.* 689 features over 3693 variants for CARMEN-E). Specially, the hyperparameters of CARMEN-E were tuned using grid searching with estimators from 700 to 1300 (step = 100), the learning rate was from 0.6 to 0.8 (step = 0.1), and the parameters of max depth and min leaf samples leaf were from 10 to 50 (step = 10); and the hyperparameters of CARMEN-F were tuned using grid searching with estimators from 700 to 1400 (step = 100), and sample weight was used in the training process. Five-fold cross-validation was employed during the model training, and the model with best area under the receiver operating characteristic curve (AUROC) over five folds was chosen and reported.

The outputs of CARMEN-E and CARMEN-F were further integrated as the "CARMEN score" which is calculated as the normalized likelihood ratio.

$$CARMEN\ score = \frac{CARMEN_E \times CARMEN_F}{CARMEN_E \times CARMEN_F + (1-CARMEN_E) \times (1-CARMEN_F)} \quad (3)$$

Two independent datasets were employed to evaluate the performance of CARMEN and nine state-of-the-art tools. The first dataset was curated from a published study on identifying cancer risk regulatory variants [35]. We took variants with adjusted $P < 0.01$ as positive cases (1164 variants) and those with adjusted $P > 0.01$ as negative cases (5375 variants). The second dataset was 43,500 eQTL- and GWAS-related variants which were tested by Biallelic Targeted STARR-seq (BiT-STARR-seq), with 2720 positive cases and 40,780 negative cases [36]. In case of data imbalance, weighted accuracy and F1 score were used as performance metrics (also see Table S5 for the detailed thresholds for the nine tools under comparison).

**Applying CARMEN to call casual variants in association studies**

For GWAS data, GWAS Catalog v1.0.2 was downloaded from https://www.ebi.ac.uk/gwas/. We also extracted the strongly linked variants ($r^2 > 0.75$) with reported lead SNPs based on the haplotypes generated by 1000 Genomes phase 3 across five populations: Utah residents (CEPH) with Northern and Western European ancestry (CEU), Han Chinese in Beijing, China (CHB), Puerto Rican in Puerto Rico (PUR), Toscani in Italia (TSI), and Yoruba in Ibadan, Nigeria (YRI). Then, CARMEN was applied to these variants to identify the potential causal variants other than the reported lead SNPs. For eQTL data, GTEx v7 multitissue variant data were obtained from https://storage.googleapis.com/gtex_analysis_v7/multi_tissue_eqtl_data/GTEx_Analysis_v7.metasoft.txt.gz. For each gene, the SNP with the smallest *P* value calculated by the RE2 model was taken as the lead SNP. Strong linkage SNPs ($r^2 > 0.75$) with each lead SNP in CEU population were further extracted. We then run CARMEN over both the reported lead SNPs and strongly linked ones.

*Luciferase reporter assay*

To further validate the prediction of CARMEN, we checked all lead SNPs with negative CARMEN scores (*i.e.*, CARMEN takes these SNPs as non-functional expression-modulating variants) in diabetes GWAS data, and pinpointed two cases with the top 2 CARMEN scores at the linked variants for follow-up luciferase assays.

We run luciferase reporter assay in HEK293T cell line which has been used by several diabetes studies [37–39]. HEK293T cells were routinely cultured in DMEM supplemented with 10% fetal bovine serum (FBS) until transfection. Cells were plated in each well of a 6-well plate and transfected at ~ 70% confluency with polyethylenimine (PEI) and 1 μg of pGL4.23 firefly vectors containing the selected fragments, using standard restriction-enzyme cloning and Renilla plasmids as a transfection control at a 1:1 ratio. Twenty-four hours after transfection, the cells were washed twice with cold 1× PBS, and then both firefly and Renilla luciferase activities were measured using the Promega's dual-luciferase reporter assay system according to the manufacturer's protocol. All luciferase activity measurements were performed in triplicate for each condition with three independent experimental replicates. Student's *t*-test was applied to estimate the statistical significance of the difference in luciferase activity between the two alleles.

## Results

**Design of CARMEN as a novel algorithm for predicting the effects of non-coding variants on both gene expression and disease risk**

Multiple studies demonstrate that non-coding variants can modulate gene expression via several distinct mechanisms [5,6,40–42]. In CARMEN annotation component, we generated 2424 features to annotate each inputted non-coding variant (Figure S3).

Briefly, we first employed CNNs to assess variants' effects on chromatin profiles based on large-scale chromatin-profiling data generated by ENCODE directly, including 1249 TF binding profiles for 596 distinct TFs, 280 DHS sites, 766 histone markers, and 108 DNA methylation profiles assessed in 36 cell lines and tissues (Figure 1B; see Method for more details as well as Table S1). Networks were trained and tuned for each profile separately, resulting in 2403 highly

accurate networks with an average AUROC of 0.908 and an average area under the precision recall curve (AUPRC) of 0.904 (Figure S5A; Table S6). Compared with DeepSEA [22], CARMEN not only covered more than twice as many features (Figure S5B) but also showed superior accuracy (Figure S6A and B, median AUROC 0.973 *vs.* 0.935, single-tailed Wilcoxon-test $P = 1.747 \times 10^{-25}$; median AUPRC 0.975 *vs.* 0.357, single-tailed Wilcoxon-test $P = 5.256 \times 10^{-154}$). Meanwhile, DNA physicochemical properties could affect gene expression by influencing the DNA shape of flanking sequences near TF binding sites [43], and we incorporated 13 DNA physicochemical properties [33] to evaluate the effects of non-coding variants on DNA physicochemical changes. The last but not the least, evolutionary conservation scores at the variants' loci derived from multiple genome alignments in the primate, mammal, and vertebrate clades were added into the annotation collection [44,45].

To reduce the risk of overfitting and high time costs of model training [46,47], we adopted a data-driven feature selection approach before training CARMEN prediction components (Figure 1C). In brief, we pretrained a multilayer neural network and estimated the feature contributions by comparing the difference in the activation value of each neuron to its 'reference activation' which was described in DeepLIFT [48]. Using the distribution of contribution scores from each dataset, we selected the features with absolute contribution scores greater than the threshold (Figure S4) for follow-up training (see Method).

To evaluate the gene expression-modulating potential of a given variant, we trained a dedicated component, CARMEN-E, based on a manually curated MPRA dataset (see Method). This component was trained with 689 feature choices from the data-driven feature selection approach (cross-validation AUROC = 0.783). Additionally, a separated component, CARMEN-F, was trained on the HGMD dataset with disease-causing regulatory variants to identify disease-causing variants with high accuracy (cross-validation AUROC = 0.921). To characterize the disease-causing variants that function through modulating gene expression, the outputs of CARMEN-E and CARMEN-F were further integrated as the CARMEN score [see Equation (3) in Method], which was used in the following evaluations.

## CARMEN shows superior performance on both large-scale independent datasets and experimentally-characterized loci compared with state-of-the-art tools

To validate the robustness of CARMEN, we first tested it on two independent datasets generated by STARR-seq. The cancer-risk dataset [35] consists of 1164 curated regulatory positive and 5375 control variants (see Method for more details on data curation; Figure S7A). Given its significantly unbalanced nature, we employed F1 score and weighted accuracy for performance comparison (**Figure 2**A). CARMEN showed significantly superior performance than the state-of-the-art tools [9,10,12,21–23,27,49,50]. Further inspection of allele-specific expression variants showed that CARMEN successfully called 70.8% (46 out of 65) of the significant allele-specific expression-modulating variants (Figure 2B, Figure S7B–D).

We further compared the performance on a curated list of experimentally-characterized variants ($n = 24$) validated by various low-throughput technologies from the literature [26,51–62], as well as on an independent luciferase-validated dataset ($n = 14$) [63]. Taken together, 28 of the 38 experimentally-validated variants were correctly reported as positive by CARMEN, showing the highest sensitivity among others (Figure 2C and D; Table S7).

Notably, benefitting from the extensive annotation generated, CARMEN is able to provide hints on the plausible mechanisms for how the variant changes the gene expression. For example, the variant rs883868 has been shown to modulate the expression of *UBASH3A* by disrupting the binding of the TF YY1 [60], which is consistent with the output of CARMEN (**Figure 3**A). Meanwhile, CARMEN annotations further showed that in 11 cell lines the alternative allele C decreases the binding affinity of YY1 (Figure 3B–D), which is further confirmed by a experimental study [60].

## CARMEN can pinpoint causal variants other than the GWAS-reported lead SNPs

Large-scale association studies, such as GWAS and eQTL studies, have identified a number of genetic variants associated with complex human diseases and traits. However, a gap between the association of a locus and the causal variant still exists, because many inherited variants and sentinel variants are in strong linkage disequilibrium (LD) regions [3,7,64]. We applied CARMEN to 51,878 reported lead SNPs extracted from the GWAS Catalog (https://www.ebi.ac.uk/gwas/home), and found that variants reported by multiple studies obtained significantly higher CARMEN scores than those that were not (single-tailed Wilcoxon-test $P = 4.233 \times 10^{-39}$), confirming previous observations [10].

To pinpoint potential causal variants other than lead SNPs, we ran CARMEN on both the reported lead SNPs and variants with strong LD ($r^2 \geq 0.75$), and we found that 45.33% of the reported lead SNPs showed significantly weaker regulatory potential than nearby variants within the same LD block ($r^2 > 0.75$). Further inspection showed that up to 60% putative casual variants identified by CARMEN were also called by at least one other tools, suggesting a potentially complementary output among different methods (Figure S8A). While most of the differences were modest, 6.65% of the variants showed differences greater than thirty-fold (Figure S8B). For example, four of them were predicted as causal variants by CARMEN and were validated in other published studies [23,57,59,65] (Figure S9). The four cases included variants associated with diseases such as inflammatory bowel disease, Behcet's disease, bladder cancer, and body mass index traits.

We next evaluated the variant rs1701704 which has been reported to be associated with susceptibility to type 1 diabetes ($P < 5 \times 10^{-8}$) by several GWAS66 [66–68], but CARMEN found that this variant showed only very weak regulatory potential (CARMEN score = 0.0024). Moreover, CARMEN pinpointed a nearby variant rs705698 with rather high potential (CARMEN score = 0.4070, **Figure 4**A). Although rs705698 has few annotations in the UCSC Genome Browser [69] (Figure S10A and B), the variant falls into the first intron of the *RAB5B* gene, a candidate gene for type 1

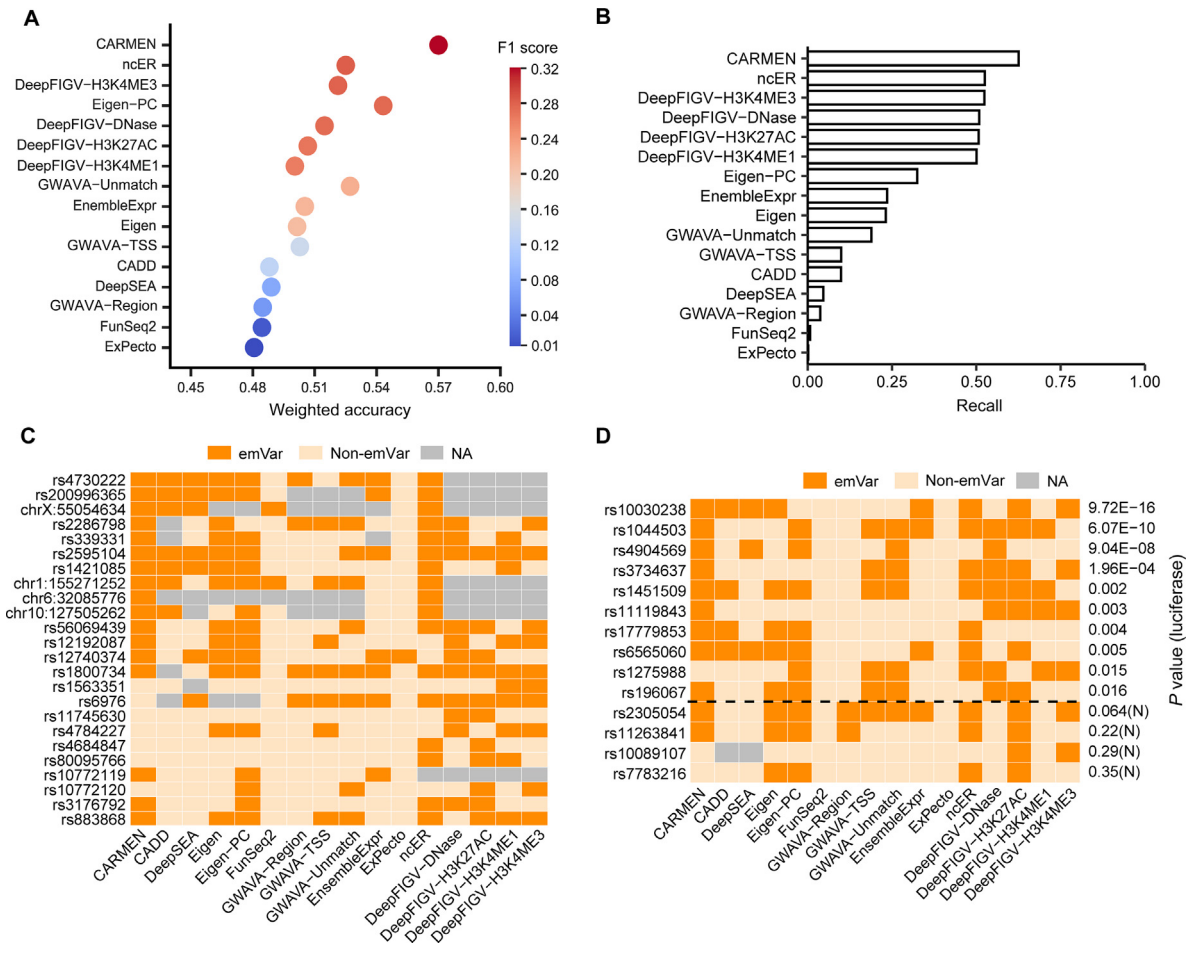


**Figure 2 The evaluation of CARMEN model performance**

**A.** Performance comparison of CARMEN to other state-of-the-art tools. The bubbles were colored by the F1 score. The thresholds of different tools were obtained from respective official websites or reference papers (Table S5). The threshold of CARMEN was evaluated on the AUROC on this dataset (Figure S7A), and the same threshold was used in follow-up figures as well as the default cutoff of the web server. **B.** Recall of different tools on the known expression-modulating variants. The x-axis indicates the proportion of variants recalled from the significant expression-modulating variants that were predicted as positive variants by the different tools. **C.** Performance comparison on a curated list of experimentally-characterized variants validated by various low-throughput technologies from the literature (Table S7). **D.** Performance comparison on an independent luciferase-validated dataset (Table S7). The right y-axis represents the reported *P* values in the luciferase assays; the four below the dotted line are the negative variants. Of note, the birth-weight-associated variant rs11119843 is missed by most of other tools. AUROC, area under the receiver operating characteristic curve; emVar, expression-modulating variant; non-emVar, non-functional expression-modulating variant; NA, not available.

diabetes [67] with strong LD ($r^2$ = 0.90746), and the annotation component of CARMEN suggested that this locus was conserved with YY1 changed in the K562, HepG2, and GM12892 cell lines, YY2 changed in the HEK293 cell line, and CBX5 and CBX1 changed in the K562 cell line (Figure S11). Independent luciferase assays further confirmed the significant change in reporter expression for rs705698 but not for rs1701704 (Figure 4B; Table S8). Notably, although CARMEN presented clear contrast on the lead and causal SNPs, many state-of-the-art tools missed them (Table S9). Likewise, variant rs1727313 has been reported as an SNP associated with type 2 diabetes ($P = 1 \times 10^{-8}$) [70]. CARMEN found no regulatory potential for this variant (CARMEN score = 0), but the linkage variant rs146239222 showed high regulatory potential (CARMEN score = 0.1999, Figure 4C), which is consistent with our luciferase reporter assay (Figure 4D; Table S8). Interestingly, variant rs146239222 is found in an enhancer region [71,72] with high H3K27ac modification, DNase clusters, and multiple TF binding sites, which is also consistent with the output of the CARMEN annotation component (Figure S10C and D) and is associated with the expression of *MPHOSPH9* in GTEx ($P = 1.34 \times 10^{-91}$, ENSG00000051825.10) [73].

Along with GWAS, eQTLs explain the variant effects on gene expression at the mRNA level [74]. When we applied CARMEN to all 7,627,599 multitissue *cis*-eQTLs reported

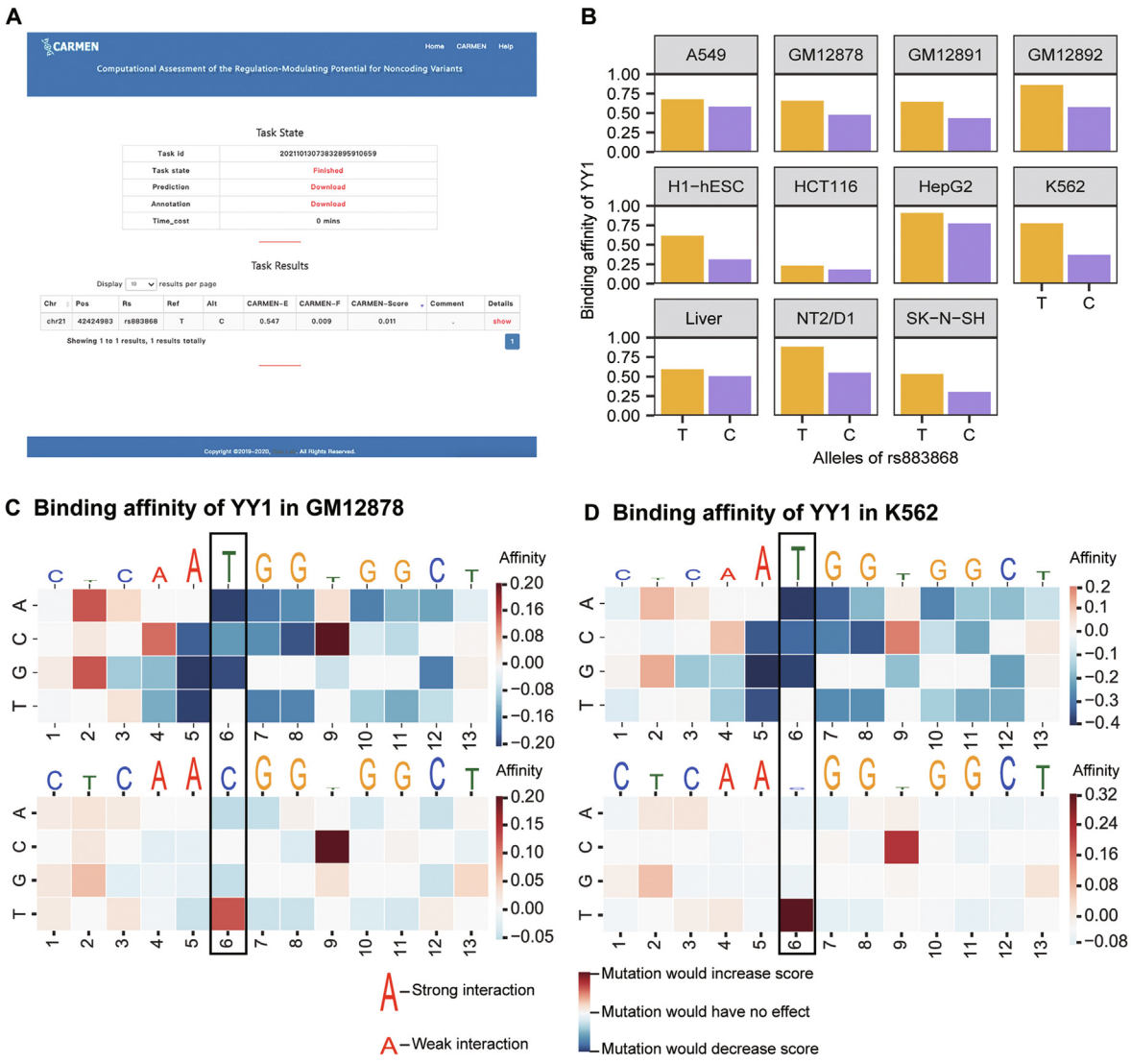


**Figure 3 CARMEN helps pinpoint the functional mechanism of non-coding expression-modulating variants**

**A.** CARMEN predicted rs883868 as an expression-modulating variant. The result was further validated by an independent CRISPR/Cas9 experiment [60]. **B.** The CARMEN annotation indicated that the alternative allele C decreases the binding affinity of YY1 in 11 cell lines. The result was also validated via 3C-qPCR [60]. **C.** Mutation maps for the variant effects on the binding affinity of YY1 in GM12878. **D.** Mutation maps for the variant effects on the binding affinity of YY1 in K562. The heatmaps indicate the changes of binding affinity of original sequence and mutated sequence. The position of variant rs883868 was displayed in the black block. 3C-qPCR, quantitative analysis of chromosome conformation capture assay.

by GTEx v7 [75], we also found several cases where the reported lead SNPs showed significantly weak regulation potential compared with the linked variants (Table S10). For example, the *SIK2*-correlated variant rs1784782 showed very weak regulatory potential (CARMEN score = 0.0004), while the linked variant rs59921976 presented rather high potential (CARMEN score = 0.6091). The independent MPRA assay [76] confirmed that the CARMEN-predicted causal variant rs59921976 had significant expression changes (two-sided Wilcoxon test, $P = 0.0147$), but the reported lead SNP rs1784782 did not.

## Discussion

Most disease-associated variants fall into non-coding regulatory elements, but their function through gene expression has not been tested [77,78]. Due to the complex mechanism of gene expression regulation, prediction, and interpretation, the function of non-coding variants remains a challenge [79]. Here, we present CARMEN to identify functional non-coding expression-modulating variants for a large-scale genomic dataset.

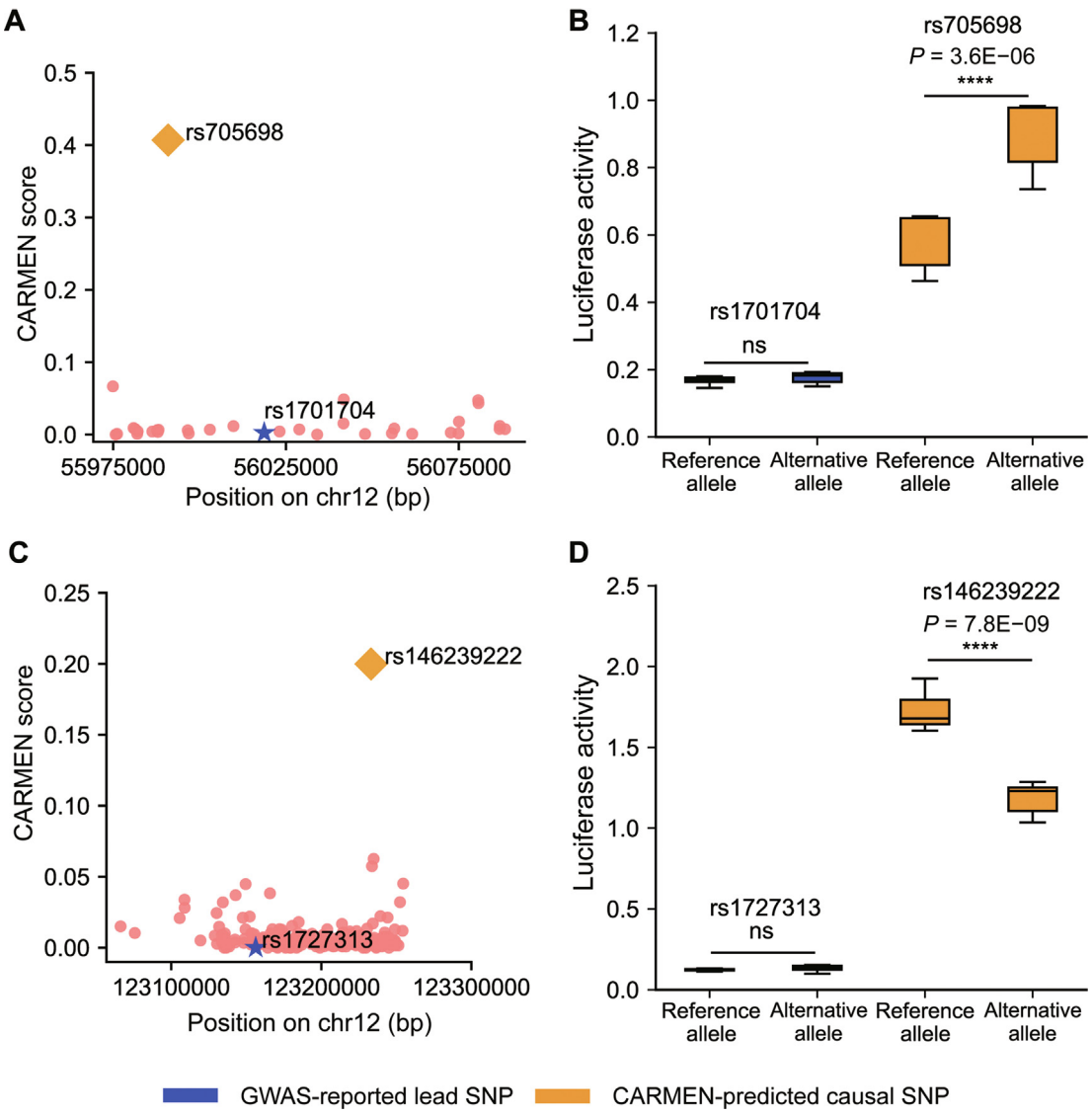


**Figure 4 Prioritizing causal variants over linkage variants with lead SNPs in GWAS**

Scatter plots show the CARMEN scores of the lead SNP and nearby variants associated with type 1 diabetes (**A**) and type 2 diabetes (**C**), respectively. The x-axis represents the genome location of each variant; the y-axis represents the CARMEN score. The blue star represents the lead SNP, which has been reported in the GWAS Catalog; the pink dots represent the variants with an LD ($r^2$) greater than 0.75 (CEU) with the lead SNP. The yellow diamond represents the causal variant with the best CARMEN score. Luciferase reporter assays validate the prediction in type 1 diabetes (**B**) and type 2 diabetes (**D**). The results were derived from more than 8 technical replicates with 3 independent experimental replications. ****, $P < 0.0001$; ns, not significant (two-sided *t*-test). The middle line of the box plot represents the median value; the box extends from first quartile to third quartile, and the whiskers (shown in black) extend to the maximum and minimum values. SNP, single-nucleotide polymorphism; GWAS, genome-wide association study; LD, linkage disequilibrium; CEU, Utah residents (CEPH) with Northern and Western European ancestry.

Inspired by pioneering works [22,28], we introduced a sequence-based annotation component for each input variant, covering multiple TF binding profiles, histone markers, DNA methylation profiles, and DHS sites across multiple cell types. Instead of one holistic network for all features [22,28], we trained independent networks for each chromatin feature separately (Figure S12A). The observed large inter-model disparity in the resulting network parameters suggested a distinct "code" for different features and supported the idea of a feature-specific network in the CARMEN annotation component (Figure S12B). Moreover, the recent rapid increase in high-throughput data also enabled more abundant training data for the CARMEN annotation component ($P = 8.05E–05$, single-tailed Wilcoxon test for 466 common features between the CARMEN annotation component and DeepSEA), further contributing to the improved model

performance. In particular, the CARMEN annotation component employed one convolutional layer architecture rather than the more complicated architecture in DeepSEA with three convolutional layers, effectively reducing the number of required parameters and enabling a more trainable and robust model for a given training set. Finally, it should be noted that DeepSEA was trained (and evaluated) based on a highly biased dataset (average positive:negative = 1:42.77, data downloaded from http://deepsea.princeton.edu/media/code/deepsea_train_bundle.v0.9.tar.gz), which might also result in poor AUPRC performance in a more realistic and balanced scheme.

Inspection over feature importance plot showed that evolutionary conservation-related features are among the mostly important ones for both CARMEN-E and CARMEN-F (Tables S11 and S12). The cumulative importance of each feature category further suggests that TF binding profiles play major roles in the CARMEN-E component, while histone markers are more "important" for CARMEN-F, suggesting that different aspects of the two component models (Figure S13).

Spatiotemporal expression specificity is important during systematic decoding of the regulatory code, and a recent analysis has highlighted that multiple genetic associations are tissue-specific [80]. One limitation for current CARMEN implementation is that it incorporates a number of multi-omics datasets generated across multiple tissues or developmental stages without explicitly modeling expression specificity, which may further introduce location-related accuracy bias (in fact, we found that CARMEN showed better performance on variants that were covered by more tissues and samples in the original ENCODE dataset, $P = 3.92 \times 10^{-114}$, single-tailed Wilcoxon test, Figure S14A). Moreover, several existing tissue-specific tools, such as ExPecto and DeepFIGV, did not perform well, even for variants with spatiotemporal expression specificity (see Figure S7B for a particular case). We believe that with the rapid development of single-cell omics [81], dynamic gene expression regulation profiling [82], and a more realistic computational model, CARMEN could be further improved.

CARMEN scales well with large-scale inputs (Figure S15) and is available as both a web server and a standalone package. Designed as a one-stop portal, the CARMEN web server supports not only on-the-fly prediction but also a user-friendly interface for visualizing the results (Figure S16). To help identify the functional mechanism of non-coding variants' regulation-modulating effects, the web server provides statistics and specific information about these features in the "Task Results" box as "show details" to help users understand the prediction results.

## Code availability

All source codes and data are available freely for academic usage at http://carmen.gao-lab.org or https://ngdc.cncb.ac.cn/biocode/tools/BT007274. The scripts for generating all figures are available at https://github.com/gao-lab/CARMEN-Figures_and_Tables, and a full package with all necessary data incorporated can be downloaded via http://carmen.gao-lab.org/download/CARMEN-results-source-codes.tar.gz.

## Competing interests

The authors declare no competing interests.

## CRediT authorship contribution statement

**Fang-Yuan Shi:** Methodology, Software, Data curation, Visualization, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Yu Wang:** Methodology, Software, Data curation, Formal analysis, Visualization. **Dong Huang:** Validation, Writing – original draft. **Yu Liang:** Data curation. **Nan Liang:** Validation. **Xiao-Wei Chen:** Validation, Investigation. **Ge Gao:** Conceptualization, Project administration, Supervision, Funding acquisition, Resources, Writing – review & editing. All authors have read and approved the final manuscript.

## Acknowledgments

We thank Drs. Zemin Zhang, Cheng Li, Letian Tao, Jian Lu, and Liping Wei at Peking University for their helpful comments and suggestions during the study. This work was supported by funds from the National Key R&D Program of China (Grant No. 2016YFC0901603), and the National High-tech R&D Program of China (Grant No. 2015AA020108), as well as the State Key Laboratory of Protein and Plant Gene Research and the Beijing Advanced Innovation Center for Genomics (ICG) at Peking University. The research of Ge Gao was supported in part by the National Program for Support of Top-notch Young Professionals. The analysis was also performed on the Computing Platform of the Center for Life Sciences of Peking University and was supported by the High-performance Computing Platform of Peking University.

## Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2021.10.003.

## ORCID

ORCID 0000-0003-4185-8129 (Fang-Yuan Shi)
ORCID 0000-0003-2799-0369 (Yu Wang)
ORCID 0000-0002-9448-4397 (Dong Huang)
ORCID 0000-0003-0695-2304 (Yu Liang)
ORCID 0000-0002-7230-245X (Nan Liang)
ORCID 0000-0001-6466-6248 (Xiao-Wei Chen)
ORCID 0000-0001-6470-8815 (Ge Gao)

## References

[1] Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annu Rev Med 2012;63:35–61.
[2] Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet 2015;24:R102–10.

[3] Gallagher MD, Chen-Plotkin AS. The post-GWAS era: from association to function. Am J Hum Genet 2018;102:717–30.
[4] Hrdlickova B, de Almeida RC, Borek Z, Withoff S. Genetic variation in the non-coding genome: involvement of micro-RNAs and long non-coding RNAs in disease. Biochim Biophys Acta 2014;1842:1910–22.
[5] Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet 2015;16:197–212.
[6] Ko YA, Yi H, Qiu C, Huang S, Park J, Ledo N, et al. Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. Am J Hum Genet 2017;100:940–53.
[7] Freedman ML, Monteiro ANA, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011;43:513–8.
[8] Edwards S, Beesley J, French J, Dunning A. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet 2013;93:779–97.
[9] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 2016;48:214–20.
[10] Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods 2014;11:294–6.
[11] 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.
[12] Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 2014;15:480.
[13] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015;31:761–3.
[14] Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep 2015;5:10576.
[15] Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 2015;47:955–61.
[16] Chen L, Jin P, Qin ZS. DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. Genome Biol 2016;17:252.
[17] Zhou L, Zhao F. Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med 2018;10:53.
[18] Bodea CA, Mitchell AA, Bloemendal A, Day-Williams AG, Runz H, Sunyaev SR. PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. Genome Biol 2018;19:173.
[19] Tivive FHC, Bouzerdoum A. A face detection system using shunting inhibitory convolutional neural networks. 2004 IEEE Int Jt Conf Neural Networks 2004;4:2571–5.
[20] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831–8.
[21] Hoffman GE, Bendl J, Girdhar K, Schadt EE, Roussos P. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. Nucleic Acids Res 2019;47:10597–611.
[22] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931–4.
[23] 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.
[24] Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics 2015;106:159–64.
[25] 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.
[26] Wakabayashi A, Ulirsch JC, Ludwig LS, Fiorini C, Yasuda M, Choudhuri A, et al. Insight into GATA1 transcriptional activity through interrogation of *cis* elements disrupted in human erythroid disorders. Proc Natl Acad Sci U S A 2016;113:4434–9.
[27] Zeng H, Edwards MD, Guo Y, Gifford DK. Accurate eQTL prioritization with an ensemble-based framework. Hum Mutat 2017;38:1259–65.
[28] Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 2016;26:990–9.
[29] Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol 2017;18:67.
[30] Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang Xi, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2020;48:D87–92.
[31] Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 2006;34:D108–10.
[32] Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. Quantifying similarity between motifs. Genome Biol 2007;8:R24.
[33] Li R, Zhong D, Liu R, Lv H, Zhang X, Liu J, et al. A novel method for *in silico* identification of regulatory SNPs in human genome. J Theor Biol 2017;415:84–9.
[34] Bishop EP, Rohs R, Parker SCJ, West SM, Liu P, Mann RS, et al. Map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. ACS Chem Biol 2011;6:1314–20.
[35] Liu S, Liu Y, Zhang Q, Wu J, Liang J, Yu S, et al. Systematic identification of regulatory variants associated with cancer risk. Genome Biol 2017;18:194.
[36] Kalita CA, Brown CD, Freiman A, Isherwood J, Wen X, Pique-Regi R, et al. High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. Genome Res 2018;28:1701–8.
[37] Small KS, Todorčević M, Civelek M, El-Sayed Moustafa JS, Wang X, Simon MM, et al. Regulatory variants at *KLF14* influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. Nat Genet 2018;50:572–80.
[38] Rusu V, Hoch E, Mercader JM, Tenen DE, Gymrek M, Hartigan CR, et al. Type 2 diabetes variants disrupt function of SLC16A11 through two distinct mechanisms. Cell 2017;170:199–212.e20.
[39] Leprêtre F, Vasseur F, Vaxillaire M, Scherer PE, Ali S, Linton K, et al. A CD36 nonsense mutation associated with insulin resistance and familial type 2 diabetes. Hum Mutat 2004;24:104.
[40] Farashi S, Kryza T, Clements J, Batra J. Post-GWAS in prostate cancer: from genetic association to biological contribution. Nat Rev Cancer 2019;19:46–59.
[41] Grossman SR, Zhang X, Wang Li, Engreitz J, Melnikov A, Rogov P, et al. Systematic dissection of genomic features determining transcription factor binding and enhancer function. Proc Natl Acad Sci U S A 2017;114:E1291–300.

[42] Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res 2011;21:381–95.
[43] Bothe M, Einfeldt E, Borschiwer M, Benner P, Vingron M, et al. Synthetic STARR-seq reveals how DNA shape and sequence modulate transcriptional output and noise. PLoS Genet 2018;14:e1007793.
[44] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110–21.
[45] 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.
[46] Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Brief Bioinform 2008;9:392–403.
[47] Ying X. An overview of overfitting and its solutions. J Phys Conf Ser 2019;1168:022022.
[48] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of 34th International Conference on Machine Learning 2017;70:3145–53.
[49] Wells A, Heckerman D, Torkamani A, Yin Li, Sebat J, Ren B, et al. Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat Commun 2019;10:5241.
[50] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47:D886–94.
[51] Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. Nature 2010;466:714–9.
[52] Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. Nature 2017;550:239–43.
[53] Li G, Cunin P, Wu D, Diogo D, Yang Y, Okada Y, et al. The rheumatoid arthritis risk variant CCR6DNP regulates *CCR6* via PARP-1. PLoS Genet 2016;12:e1006292.
[54] Miller CL, Anderson DR, Kundu RK, Raiesdana A, Nürnberg ST, Diaz R, et al. Disease-related growth factor and embryonic signaling pathways modulate an enhancer of *TCF21* expression at the 6q23.2 coronary heart disease locus. PLoS Genet 2013;9:e1003652.
[55] Spisák S, Lawrenson K, Fu Y, Csabai I, Cottman RT, Seo JH, et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med 2015;21:1357–63.
[56] Cowper-Sal·lari R, Zhang X, Wright JB, Bailey SD, Cole MD, Eeckhoute J, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for *FOXA1* and alter gene expression. Nat Genet 2012;44:1191–8.
[57] Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO obesity variant circuitry and adipocyte browning in humans. N Engl J Med 2015;373:895–907.
[58] Lawrenson K, Kar S, McCue K, Kuchenbaeker K, Michailidou K, Tyrer J, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. Nat Commun 2016;7:12675.
[59] Pattison JM, Posternak V, Cole MD. Transcription factor KLF5 binds a *Cyclin E1* polymorphic intronic enhancer to confer increased bladder cancer risk. Mol Cancer Res 2016;14:1078–86.
[60] Gao P, Uzun Y, He B, Salamati SE, Coffey JKM, Tsalikian E, et al. Risk variants disrupting enhancers of $T_H1$ and $T_{REG}$ cells in type 1 diabetes. Proc Natl Acad Sci U S A 2019;116:7581–90.
[61] Klein JC, Keith A, Rice SJ, Shepherd C, Agarwal V, Loughlin J, et al. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. Nat Commun 2019;10:2434.
[62] Claussnitzer M, Dankel S, Klocke B, Grallert H, Glunk V, Berulava T, et al. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. Cell 2014;156:343–58.
[63] Wang X, Tucker NR, Rizki G, Mills R, Krijger PHL, de Wit E, et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. Elife 2016;5:e10557.
[64] Wray NR, Purcell SM, Visscher PM, Flint J. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol 2011;9:e1000579.
[65] Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, Sun J, et al. A prostate cancer susceptibility allele at 6q22 increases *RFX6* expression by modulating *HOXB13* chromatin binding. Nat Genet 2014;46:126–35.
[66] Petukhova L, Duvic M, Hordinsky M, Norris D, Price V, Shimomura Y, et al. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. Nature 2010;466:113–7.
[67] Hakonarson H, Qu HQ, Bradfield JP, Marchand L, Kim CE, Glessner JT, et al. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. Diabetes 2008;57:1143–6.
[68] Plagnol V, Howson JMM, Smyth DJ, Walker N, Hafler JP, Wallace C, et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet 2011;7:e1002216.
[69] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006.
[70] Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 2014;46:234–44.
[71] Kang R, Zhang Y, Huang Q, Meng J, Ding R, Chang Y, et al. EnhancerDB: a resource of transcriptional regulation in the context of enhancers. Database 2019;2019:bay141.
[72] de Rie D, Abugessaisa I, Alam T, Arner E, Arner P, Ashoor H, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol 2017;35:872–8.
[73] Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. Biopreserv Biobank 2015;13:311–9.
[74] Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60.
[75] Duong D, Gai L, Snir S, Kang EY, Han B, Sul JH, et al. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. Bioinformatics 2017;33:i67–74.
[76] van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, et al. High-throughput identification of human SNPs affecting regulatory element activity. Nat Genet 2019;51:1160–9.
[77] Farh KH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337–43.
[78] 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.

[79] Zhu Y, Tazearslan C, Suh Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. Exp Biol Med 2017;242:1325–34.

[80] Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 2018;9:1825.

[81] Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nat Commun 2020;11:810.

[82] Strober BJ, Elorbany R, Rhodes K, Krishnan N, Tayeb K, Battle A, et al. Dynamic genetic regulation of gene expression during cellular differentiation. Science 2019;364:1287–90.