Computational Assessment of the Regulation-Modulating Potential for Noncoding Variants

Fang-Yuan Shi¹, Yu Wang¹, Dong Huang², Yu Liang³, Nan Liang¹, Xiao-Wei
Chen^{2,4}, Ge Gao^{1*}

¹ Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation
Center for Genomics (ICG), Center for Bioinformatics (CBI), and State Key
Laboratory of Protein and Plant Gene Research at School of Life Sciences, Peking
University, Beijing, 100871, China.

² The 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, China.

⁴ Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary
Studies, Peking University, Beijing, 100871, China

Email:

FYS: shify@mail.cbi.pku.edu.cn

YW: wangy@mail.cbi.pku.edu.cn

DH: huangdong@pku.edu.cn

YL: liangyu@email.ncu.edu.cn

NL: liangn@mail.cbi.pku.edu.cn

XWC: xiaowei_chen@pku.edu.cn

24 GG: gaog@mail.cbi.pku.edu.cn

25

26 **Abstract**

27 Large-scale genome-wide association and expression quantitative trait loci studies have identified
 28 multiple noncoding variants associated with genetic diseases via affecting gene expression.
 29 However, effectively and efficiently pinpointing causal variants remains a serious challenge. Here,
 30 we developed CARMEN, a novel algorithm to identify functional noncoding expression-
 31 modulating variants. Multiple evaluations demonstrated CARMEN's superior performance over
 32 state-of-the-art tools. Its higher sensitivity and low false discovery rate enable CARMEN to identify
 33 multiple causal expression-modulating variants that other tools simply missed. Meanwhile,
 34 benefitting from extensive annotations generated, CARMEN provides mechanism hints on
 35 predicted expression-modulating variants, enabling effectively characterizing functional variants
 36 involved in gene expression and disease-related phenotypes. CARMEN scales well with the
 37 massive datasets and is available online as a Web server at <http://carmen.gao-lab.org>.

38

39 **Keywords**

40 Noncoding variants, Expression-modulating variants, Gene regulation, Web server

41

Approximately 98% of the human genome is noncoding¹. While genome-wide association studies (GWAS) have identified thousands of disease-associated or complex trait-associated variants^{2,3} and more than 90% of disease-associated variants are located in noncoding regions⁴, their biological functions and mechanisms remain elusive^{5,6}.

Several tools have been proposed to prioritize functional noncoding variants based on existing annotations at corresponding loci^{7–10}. Recently, convolutional neural networks (CNNs)¹¹ have been widely used to characterize the regulatory activity of sequences, enabling the calling of regulatory variants that result in significant changes in binding affinity^{12,13} based on sequence only.

The massively parallel reporter assay (MPRA) enables the systematic screening of potential regulatory variants and provides valuable candidate training datasets to generate *in silico* prediction tools that predict the variants' effects on gene expression. EnsembleExpr¹⁴ uses MPRA data to predict the causal variants in eQTLs, and ExPecto¹⁵ *ab initio* predicts the variants' effects on gene expression from 40-kb promoter-proximal sequences based on reference data but not mutagenesis data.

Here, we developed CARMEN, an algorithm framework for predicting the effects of noncoding variants on both gene expression and disease risk (Fig. 1a). Compared with state-of-the-art tools, CARMEN shows superior performance on both high-throughput datasets and low-throughput case studies and has higher sensitivity with low false discovery rate, enabling effective identification of multiple causal expression-modulating variants that other tools missed. We also illustrated that CARMEN can pinpoint causal variants other than the reported lead SNPs in GWAS and eQTLs. CARMEN is available as a Web server with free access for academic usage at <http://carmen.gao-lab.org>.

Results

Multiple studies demonstrate that noncoding variants can modulate gene expression via several distinct mechanisms^{16–18}. Both TF binding and epigenetic modifications contribute to gene regulation^{19,20}. Inspired by DeepSEA¹², we trained CNN models with large-scale chromatin-profiling data generated by Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomic Project directly, including 1,249 TF binding profiles for 596 distinct TFs, 280 DHS profiles, 766 histone-mark profiles and 108 DNA methylation profiles assessed in 36 cell lines and tissues (see Methods for more details as well as Supplementary Fig. 1 and Supplementary Table 1). Networks were trained and tuned for each profile separately, resulting in 2,403 highly accurate models with an average AUROC (Area Under the Receiver Operating Characteristic Curve) of 0.908 and an average AUPRC (Area Under the Precision Recall Curve) of 0.904 (Supplementary Fig. 2b and Supplementary Table 2). Compared to DeepSEA, CARMEN not only covers more than twice as many features (Supplementary Fig. 2a) but also shows superior accuracy (Supplementary Fig. 3, 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 $p = 5.256 \times 10^{-154}$).

DNA physicochemical properties also affect gene expression through influencing the shape properties of flanking sequences near transcription factor binding sites²¹. We incorporated 13 conformational and physicochemical DNA properties²² to evaluate the noncoding variants' effects on DNA shape change.

Finally, along with the eight PhyloP²³ and PhastCons²⁴ conservation scores derived from multiple genome alignments in the primate/mammal/vertebrate clades, CARMEN generated 2,424 candidate features for each inputted variant.

To reduce the risk of overfitting and high time costs of model training, we adopted a data-driven feature selection approach before model training. Briefly, we pretrained a multilayer neural network and estimated features contributions by comparing the difference in the activation value of each neuron to its 'reference activation'²⁵. Using the distribution of contribution scores from each dataset, we selected the features with absolute contribution scores greater than the threshold (Supplementary Fig. 6) for follow-up training (**Methods**).

To evaluate the gene expression-modulating potential of a given variant, we trained a dedicated module, CARMEN-E, based on a manually curated MPRA dataset (**Methods**). Trained with 689

features choice from the data-driven feature selection approach, the CARMEN-E module had significantly superior performance than both EnsembleExpr and ExPecto (Fig. 1b). Additionally, a separate module, CARMEN-F, was trained on the HGMD Disease-causing Mutation (DM) in a regulatory-variant dataset to identify disease-causing variants with high accuracy (cross-validation AUROC = 0.921, also see Fig. 1c for CARMEN-F evaluation on an independent pathogenic dataset).

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:

$$\text{CARMEN Score} = \frac{\text{CARMEN_E} \times \text{CARMEN_F}}{\text{CARMEN_E} \times \text{CARMEN_F} + (1 - \text{CARMEN_E}) \times (1 - \text{CARMEN_F})}$$

which will be used in the following evaluations.

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

To validate the robustness of CARMEN, we first tested it on two independent datasets generated by STARR-seq²⁶.

The cancer-risk dataset²⁶ consists of 1,164 curated regulatory positive and 5,375 control variants (see Methods for more details on data curation as well as Supplementary Fig. 7a). Given its significantly unbalanced nature, we employed the F1 score and weighted accuracy for performance comparison. CARMEN shows significantly superior performance than the state-of-the-art tools^{7,10,12,14,15,27,28}. Of note, we found that DeepSEA and ExPecto performed worse, partly due to their surprisingly low sensitivity (Fig. 1d). Further inspection of allele-specific expression (ASE) variants showed that CARMEN successfully called 69.2% (45 out of 65) of significant allele-specific expression-modulating variants, while DeepSEA and ExPecto found only 7 and 1, respectively (Fig. 1e, also see Supplementary Fig. 7b, c, d for evaluation on the additional dataset).

We further compared the performance on a curated list of experimentally characterized variants validated by various low-throughput technologies from the literature^{29,30,39–41,31–38}. Out of 24 experimentally validated variants, 17 were correctly reported as positive by CARMEN, showing the highest sensitivity among others (Fig. 1f and Supplementary Table 3).

Notably, benefitting from the extensive annotation generated, CARMEN is able to provide hints on possible mechanisms for how the variant changes the gene expression. For example, the variant rs883868 has been shown to modulate the expression of UBASH3A by disrupted binding of transcription factor YY1⁴¹, which is consistent with the output of the CARMEN annotations (Fig. 2).

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

Large-scale association studies, such as GWAS and eQTLs, 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 regions^{3,5,42}. We applied CARMEN on 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 observations²⁷.

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 found that 45.33% of the reported lead SNPs show significantly weaker regulatory potential than nearby variants within the same LD block ($r^2 > 0.75$). While most of the differences were modest, 6.65% of variants showed differences larger than thirty-fold. (Supplementary Fig. 8)

For example, several GWA studies report that variant rs1701704 is associated with the susceptibility of type 1 diabetes ($p < 5 \times 10^{-8}$)⁴³⁻⁴⁵, but CARMEN found that this variant shows only very weak regulatory potential (CARMEN score = 0.0024). Meanwhile, CARMEN pinpointed a nearby variant rs705698 with rather high potential (CARMEN score = 0.4070, Fig. 3a). While rs705698 has few annotations in the UCSC Genome Browser⁴⁶ (Supplementary Fig. 10a,b), the variant falls into the first intron of gene RAB5B, which is a candidate gene of type 1 diabetes⁴⁴ with strong linkage ($r^2=0.90746$), and the annotation module of CARMEN suggests that it is a conserved locus with YY1 changed in the K562, HepG2, GM12892 cell lines, YY2 changed HEK293 cell line, CBX5 and CBX1 changed in the K562 cell line (Supplementary Fig. 11). Independent luciferase assays further confirmed the significant change in report expression for rs705698 but not for rs1701704 (Fig. 3b and Supplementary Table 11). Notably, while CARMEN presents crystal clear contrast on the lead SNP and causal SNP, many state-of-the-art tools missed them (Fig. 3e and Supplementary Table 9). Likewise, variant rs1727313 was reported as an SNP associated with type

170 2 diabetes ($p = 1 \times 10^{-8}$)⁴⁷. CARMEN found no regulatory potential for this variant (CARMEN
171 score = 0), but the linkage variant rs146239222 showed high regulatory potential (CARMEN score
172 = 0.1999, Fig. 3c), which was consistent with our luciferase reporter assay (Fig. 3d and
173 Supplementary Table 11). Interestingly, variant rs146239222 is found in an enhancer region^{48,49}
174 with high H3K27ac modification, DNase clusters, and multiple transcription factor binding, which
175 is also consistent with the output of the CARMEN annotation module (Supplementary Fig. 10c,d),
176 and is associated with the expression of MPHOSPH9 in GTEx ($p = 1.34 \times 10^{-91}$,
177 ENSG00000051825.10)⁵⁰.

178

179

Discussion

Expression quantitative trait loci (eQTL) explain the variant effects on gene expression at the mRNA level⁵¹. When we applied CARMEN on all 7,627,599 multi-tissue *cis*-eQTLs reported by GTEx v7⁵², we also found several cases where the reported lead SNP showed significantly weak regulation potential compared with that of the linked variants (Supplementary Table 6). For example, the SIK2-correlated variant rs1784782 shows very weak regulatory potential (CARMEN score = 0.0004), while the linked variant rs59921976 presents rather high potential (CARMEN score = 0.6091). The independent MPRA assay confirmed that the CARMEN-predicted causal variant rs59921976 had significant expression changes, but the reported lead SNP rs1784782 did not⁵³, and the validated expression-modulating variant rs59921976 was also missed by ExPecto as a nonsignificant variant.

We calculated feature importance for all CARMEN-E and CARMEN-F features. Consistent with previous reports^{28,54}, evolutionary conservation is found to be important for both modules. Furthermore, transcription factor binding profiles play major roles in the CARMEN-E module, while histone markers are more “important” for CARMEN-F, suggesting different aspects of the two modules model (Supplementary Fig. 9, Supplementary Table 4 and Supplementary Table 5).

CARMEN is well tuned for large-scale data, running at ~10,000 variants per hour on a modern Linux server (Supplementary Fig. 13). CARMEN is available as both a webserver and a standalone package at <http://carmen.gao-lab.org>. 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 (Supplementary Fig. 12). To help identify the functional mechanism of noncoding 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.

Online Methods

Learning regulation code by sequence-orientated model

Predict variants' effect with sequence-orientated models

The raw data for 1,249 transcription factor (TF) binding profiles, 766 histone markers, 280 DNase I – hypersensitive sites and 108 DNA methylation profiles were downloaded from ENCODE (Supplementary Table 1). We filtered the data to ensure quality. For TF data, only those involved in conservative irreproducible discovery rate (IDR) peaks and optimal IDR peaks were included. For the data about histone marker profiles, replicated peak data were included. For DNase I – hypersensitive sites, pseudoreplicated IDR peak data were used. All data were in bed format, and the reference genome version was GRCh38. The same features (such as CTCF) from different cell lines and using different experimental treatment methods were considered as different chromatin features.

To prepare the input for the model, we took a 200 bp window centered on the ChIP-seq peak or target methylation site from all data. For some features, there were several data files from different labs or experiments, and as a result, these features were merged. The merging standard used was that if two sequences from two data files for the same feature overlapped each other, the sequence with a lower peak value was abandoned. Then, the bed files were converted to FASTA files to obtain sequences from the reference genome. During this process, the reverse complement strands were also converted (Supplementary Fig. 4).

For the classification model, positive datasets and control datasets were both needed. For ChIP-seq data, we removed the positive data from the reference genome and then split the rest into 200 bp bins. We randomly selected some bins to make the number of the positive and control dataset equal. For methylation data, if the methylation rate of the target site was higher than 50%, it was considered positive data; if the rate was lower than 50%, it was considered control data; and if the rate was 50%, it was abandoned. Positive data were labeled as “1” and control data as “0”.

Training and testing sets were split randomly in an 85% to 15% ratio. To obtain a reliable and stable model, five-fold cross-validation was conducted.

Structure of the model

The classification model was a deep convolutional network (Supplementary Fig. 5). The basic layer types in the model were a convolution layer, a pooling layer and a fully connected layer. The first layer of the model was a 1D-convolutional layer, which was responsible for detecting features along the sequences with the activation function ReLU. This layer had 128 or 256 kernels, and the length of each kernel was 4, 10, 12 or 20. Each kernel could be considered a PWM calculator that moved on the sequence with step one. Formally,

$$\text{Conv}(X)_{ik} = \text{ReLU}\left(\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W^k X_i + m, n\right)$$

241 where X is the input, i is the index of the output position and k is the index of the kernel. W^k is the
242 weight matrix whose shape was $M \times N$, M is the kernel length and N is the channel number of the
243 input. The second layer was a 1D-max-pooling layer with a pooling size of 10 or 20. It was used to
244 take the maximal value from nonoverlapping windows. The function of the max-pooling layer was
245 to reduce the dimensionality of the input data. It was defined as

$$\text{MaxPool}(X)_{ik} = \max(\{X_{iM,k}, X_{iM+1,k}, \dots, X_{iM+M-1,k}\})$$

246 where M is the pooling size.

247 The following layers were the dropout layers and fully connected layers with ReLU as the
248 activation function. The fully connected layer multiplied linear combinations of the input and was
249 defined as:

$$F(X) = W_{mn}X + b_m$$

250 where m is the output channel number and n is the input channel number. The dropout layer was
251 used to reduce overfitting by randomly setting a certain proportion of input units to zero during
252 training. We set m to 512 the drop rate to 0.5.

253 The last layer of the model was a fully connected layer with a sigmoid function as the activation
254 function. The sigmoid function scales the prediction to 0-1 and is defined as

$$\text{Sigmoid}(X) = \frac{1}{1 + e^X}$$

255 *Training of the model*

256 The loss function of the model was binary cross entropy (BCE). Specifically,

$$\text{BCE}(t, p) = -(t \times \log(p) + (1 - t) \times \log(1 - p))$$

257 where t is the true label and p is the prediction of the model.

258 The optimizer was stochastic gradient descent (SGD), and the learning rate was 0.01. Early stop
259 was used during the training process. The model ran 50 epochs for training. The classification
260 model's performance was evaluated in terms of the classification rate by using the area under the
261 receiver operating characteristic curve (AUROC).

262 We obtained a 100 bp flanking sequence upstream and a 99 bp flanking sequence downstream of
263 each allele, resulting in a 200 bp reference sequence and a 200 bp alternative sequence for each
264 variant. With each sequence-oriented model, we calculated the log2 fold changes (as the method
265 shown in DeepSEA) of the reference and alternative sequences as the variant effects
266 (Supplementary Fig. 5).

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

Implementation

Implementation of the model utilized the deep learning framework TensorFlow (<https://www.tensorflow.org/>), and all scripts for the models and evaluations were written in Python (<https://www.python.org/>). An NVIDIA Tesla P100 Graphics Processing Unit was used to train the model.

Predicting variant effects on conformational and physicochemical DNA properties

As Rong Li et al.²² showed in their paper, 13 conformational and physicochemical DNA properties are useful to find functional elements in the genome, and thus we used the method referred to by Rong Li et al. to calculate the variant effect on these 13 features.

For the 12 physicochemical features selected from DiProDB⁵⁵, we obtained a 2 bp flanking sequence on both sides of each allele, resulting in a 5 bp reference sequence and a 5 bp alternative sequence. A 2 bp-wide sliding window is used for the 5 bp sequence, and the values from each window are then summed. For the hydroxyl radical cleavage pattern (OH)⁵⁶, we obtained a 6 bp flanking sequence on both sides of each allele, resulting in a 13 bp sequence for each allele. The score was calculated using a 4 bp-wide window for the 13 bp sequence. Because the OH feature was influenced by the double strands of DNA, the OH score for each allele was the average score of the two strands.

Predicting variant effects on evolutionary conservation features

Further features include the conservation score of each position of the variants. We used hgwiggle to download 8 kinds of conservation scores: 4 PhastCons and 4 PhyloP scores from the UCSC Genome Browser in reference genome version GRCh37.

Feature selection

We annotated each variant with each feature, and we used a multilayer perceptron neural network to introduce the complex relationship of different features to improve the flexibility of feature interactions (Supplementary Fig. 6). The first three layers learn the code to represent input features by the hidden units.

$$h = \text{ReLU}(W^{(1)}X + b^{(1)})$$

$$d = \text{linear}(W^{(2)}h + b^{(2)})$$

A dense layer with L1 regularization is added behind these layers. We added a $\lambda_1 = 0.0001$ penalty to the model weight and a $\lambda_2 = 0.001$ penalty to the output of this layer to increase the sparsity of features.

$$s = \lambda_2 \left\| \lambda_1 \|W^{(3)}\|_1 d + b^{(3)} \right\|_1$$

The last two layers of the model are fully connected layers, one with a ReLU activation function and the other with a sigmoid function. The sigmoid function scales the output ranges from 0 to 1.

$$z = \text{sigmoid} \left(W^{(5)} \left(\text{ReLU} \left(W^{(4)} s + b^{(4)} \right) \right) + b^{(5)} \right)$$

We label the control variants labels as 0 and the positive variants as 1. To solve the binary classification problem of whether the variant has an effect on modulating expression, we trained the multilayer perceptron neural network using 2,424 features with a step size of 300 and then selected the best model as the basic model to perform feature selection.

To learn the important features, we used Deep Learning Important Features (DeepLIFT)²⁵ to calculate the contribution score of each variants' features 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 score of the input data and selected the features with contribution scores higher than the absolute cutoff. We selected features with contribution scores greater than 0.5 or less than -0.5 in the CARMEN-E module and contribution scores greater than 2 or less than -2 in the CARMEN-F module. With two training datasets, after feature selection, we then trained the Adaboost Decision Tree with 689 features and the Random Forest with 1,190 features.

The loss function was binary cross entropy, and we used 'Adadelta' to optimize the models with default parameters. The batch size was 100, and each model was trained with 100 epochs. We utilized the early-stopping method and set the patience to 30 to avoid overfitting. We used Keras version 1.1.0 and Python version 2.7.12 to implement the model, and NVIDIA-SMI, Driver version 381.22, with a GTX 1080 Ti to train the model.

To analyze the different types of feature importance, scikit-learn 0.19.1 was used to calculate feature importance determined by its contributions to the regulation-modulating variant classification. Then, we calculated the cumulative feature importance for each feature category.

Model selection

To optimize the hyperparameters of the model and select the best model for feature selection, we used hidden units, ranging from 1 to the number of features with a step size depending on the number of features. Random-seed for cross-validation, random-seed for models and five-fold cross-validation were used for model training. We then selected the best model after using the AUROC to evaluate the performance of the models.

Predicting the variant effects model

Dataset Curation

In CARMEN prediction, two datasets were used for model training (details in Supplementary Table 7 and Supplementary Table 8). In the expression module, the positive dataset was the mutagenesis data of experimental validated variants, because high-quality labeled variants can facilitate the methods to distinguish regulatory variants. Since 2009, the parallel reporter assay has been used to study promoters⁵⁷, and later, this assay was modified to derive several assays, such as the MPRA and STARR-seq. Here, we curated data from 4 papers that applied MPRA to confirm regulatory variants. In the function module, the positive variants were from HGMD professional version 2018.4 with disease-causing variants in regulatory mutations and the control variants from 1000 Genomes. When comparing the performance of CARMEN-F with other tools, since we randomly selected the 1000 Genomes variants, it is possible that these data were used in training dataset for other tools, and so we curated another independent pathogenic dataset (Supplementary Table 7). We excluded variants that could not map from the GRCh37 to the GRCh38 versions of the human reference genome.

CARMEN prediction consisted of two modules. The first was CARMEN-E, which was trained with an expression-related dataset to evaluate the variants' effects on gene expression. After the feature selection process, we obtained 689 features. Then, we used the Adaboost Decision Tree to train this dataset with five-fold cross-validation. The other module was CARMEN-F, which evaluated the effect of variants on disease risk. After the feature selection process, we obtained 1,190 features. Random forest with sample weights was used to train the model with a disease-related dataset using five-fold cross-validation. AUROC was used to evaluate model performance. The CARMEN score was calculated with the CARMEN-E and CARMEN-F scores with Bayes' formula.

$$\text{CARMEN Score} = \frac{\text{CARMEN_E} \times \text{CARMEN_F}}{\text{CARMEN_E} \times \text{CARMEN_F} + (1 - \text{CARMEN_E}) \times (1 - \text{CARMEN_F})}$$

Model evaluation and application

The model was evaluated with two independent datasets with a large number of variants previously tested by self-transcribing active regulatory region sequencing (STARR-seq) in 2018. The first dataset was curated from a study that systematically identified regulatory variants with cancer risk. In this study, the author identified 1,333 variants in the fragments that regulated gene expression from 10,673 variants. Filtering out the variants' reference or alternative alleles in the fragments that did not have statistical significance (DESeq2⁵⁸ *p* values) and the locus that could not map from GRCh37 to GRCh38, we obtained 6,539 variants. Among these variants, 1,164 variants in

fragments that regulated gene expression and 65 variants had different regulatory activities in two alleles.

We also curated another dataset based on the Biallelic Targeted STARR-seq⁵⁹, which tested 43,500 variants and identified 2,720 variants with significant ASEs. For these two unbalanced datasets, we calculated the weighted accuracy and F1 score to evaluate the performance. The thresholds of the tools that we compared can be found in Supplementary Table 10.

GWAS Catalog v1.0.2 was downloaded from <https://www.ebi.ac.uk/gwas/>. We also extracted the strongly linked variants (i.e. $r^2 > 0.75$) with reported lead SNP based on the haplotypes generated by 1000 Genomes phase 3 across five populations (CEU, CHB, PUR, TSI, YRI). Then, CARMEN was applied to these variants to identify the potential causal variants other than the reported lead SNP.

GTEx v7 multi-tissue variants data were from https://storage.googleapis.com/gtex_analysis_v7/multi_tissue_eqtl_data/GTEx_Analysis_v7.metasoft.txt.gz. Metasoft extends the random effects (REs) meta-analysis model into the RE2 model⁵², which was used in GTEx v7 to evaluate the SNP effects on gene expression across multiple tissues. Furthermore, we wanted to identify the variants that were labeled by the RE2 model as lead SNPs but not the causal variants for the eGenes. We used CARMEN to predict all the variants except those without GRCh38 loci or conservation scores. For each gene, the lead SNP was the most significant variant with the smallest p value calculated by the RE2 model. We also extracted linkage SNPs with an LD r^2 greater than 0.75 with each lead SNP in the CEU population. The lead SNPs in the GTEx multi-tissue eQTLs were all predicted by CARMEN. If the lead SNPs were predicted to be negative by CARMEN, then we selected variants with LD r^2 greater than 0.75 with the lead SNPs. Moreover, we required the high linkage variants were not existed in GTEx multi-tissue eQTLs. Then, we used CARMEN to pinpoint the predicted causal SNPs from the negative lead SNPs in each eGene.

Luciferase reporter assay

For the luciferase reporter assay, HEK293T cells were routinely cultured in DMEM/10% FBS until ready for transfection. Cells were plated in each well of a 6-well plate and transfect the cells at ~70% confluency with PEI and transfected with 1 μ g pGL4.23 firefly vectors containing the selected fragments using standard restriction-enzyme cloning and Renilla plasmids as a transfection control in a 1:1 ratio. Twenty-four hours after transfection, cells were washed twice with cold 1 $\times$ PBS, and firefly and Renilla luciferase activity was measured using the Dual-Luciferase Reporter

Assay System (Promega) according to the manufacturer's protocol. All the 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.

Availability of data and materials

All source codes and data are available freely for academic usage at <http://carmen.gao-lab.org>. The scripts for generating all figures could be downloaded via <http://carmen.gao-lab.org/download/CARMEN-results-source-codes.tar.gz>

References

1. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human Genome Sequencing in Health and Disease. *Annu. Rev. Med.* **63**, 35–61 (2012).
2. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).
3. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
4. Hrdlickova, B., de Almeida, R. C., 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 - Mol. Basis Dis.* **1842**, 1910–1922 (2014).
5. Freedman, M. L. et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.* **43**, 513–8 (2011).
6. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
7. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).
8. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
9. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
10. Fu, Y. et al. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
11. Tivive, F. H. C. & Bouzerdoun, A. A face detection system using shunting inhibitory convolutional neural networks. in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)* **4**, 2571–2575 (IEEE, 1998).
12. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
13. Alipanahi, B., DeLong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–8 (2015).
14. Zeng, H., Edwards, M. D., Guo, Y. & Gifford, D. K. Accurate eQTL prioritization with an ensemble-based framework. *Hum. Mutat.* **38**, 1259–1265 (2017).
15. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).

- 445 16. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease.
446 *Nat. Rev. Genet.* **16**, 197–212 (2015).
- 447 17. Ko, Y.-A. et al. Genetic-Variation-Driven Gene-Expression Changes Highlight Genes with
448 Important Functions for Kidney Disease. *Am. J. Hum. Genet.* **100**, 940–953 (2017).
- 449 18. Farashi, S., Kryza, T., Clements, J. & Batra, J. Post-GWAS in prostate cancer: from genetic
450 association to biological contribution. *Nat. Rev. Cancer* **19**, 46–59 (2019).
- 451 19. Grossman, S. R. et al. Systematic dissection of genomic features determining transcription
452 factor binding and enhancer function. *Proc. Natl. Acad. Sci.* **114**, E1291–E1300 (2017).
- 453 20. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell*
454 *Res.* **21**, 381–95 (2011).
- 455 21. Schöne, S. et al. Synthetic STARR-seq reveals how DNA shape and sequence modulate
456 transcriptional output and noise. *PLOS Genet.* **14**, e1007793 (2018).
- 457 22. Li, R. et al. A novel method for in silico identification of regulatory SNPs in human genome.
458 *J. Theor. Biol.* **415**, 84–89 (2017).
- 459 23. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral
460 substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
- 461 24. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
462 genomes. *Genome Res.* **15**, 1034–1050 (2005).
- 463 25. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through
464 Propagating Activation Differences. *arXiv* 1–9 (2017).
- 465 26. Liu, S. et al. Systematic identification of regulatory variants associated with cancer risk.
466 *Genome Biol.* **18**, 194 (2017).
- 467 27. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding
468 sequence variants. *Nat. Methods* **11**, 294–296 (2014).
- 469 28. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human
470 genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- 471 29. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol
472 locus. *Nature* **466**, 714–719 (2010).
- 473 30. Wakabayashi, A. et al. Insight into GATA1 transcriptional activity through interrogation of
474 cis elements disrupted in human erythroid disorders. *Proc. Natl. Acad. Sci.* **113**, 4434–4439
475 (2016).
- 476 31. Klein, J. C. et al. Functional testing of thousands of osteoarthritis-associated variants for
477 regulatory activity. *Nat. Commun.* **10**, 2434 (2019).
- 478 32. Claussnitzer, M. et al. Leveraging Cross-Species Transcription Factor Binding Site Patterns:
479 From Diabetes Risk Loci to Disease Mechanisms. *Cell* **156**, 343–358 (2014).

- 480 33. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–
481 243 (2017).
- 482 34. Li, G. et al. The Rheumatoid Arthritis Risk Variant CCR6DNP Regulates CCR6 via PARP-1.
483 *PLoS Genet.* **12**, e1006292 (2016).
- 484 35. Miller, C. L. et al. Disease-Related Growth Factor and Embryonic Signaling Pathways
485 Modulate an Enhancer of TCF21 Expression at the 6q23.2 Coronary Heart Disease Locus.
486 *PLoS Genet.* **9**, e1003652 (2013).
- 487 36. ndor Spis k, S. et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing
488 function of noncoding GWAS variants. *Nat. Med.* **21**, 1357–1363 (2015).
- 489 37. Cowper-Sallari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of
490 chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
- 491 38. Claussnitzer, M. et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.
492 *N. Engl. J. Med.* **373**, 895–907 (2015).
- 493 39. Lawrenson, K. et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1
494 breast-ovarian cancer susceptibility locus. *Nat. Commun.* **7**, 12675 (2016).
- 495 40. Pattison, J. M., Posternak, V. & Cole, M. D. Transcription Factor KLF5 Binds a Cyclin E1
496 Polymorphic Intronic Enhancer to Confer Increased Bladder Cancer Risk. *Mol. Cancer Res.*
497 **14**, 1078–1086 (2016).
- 498 41. Gao, P. et al. Risk variants disrupting enhancers of TH1 and TREG cells in type 1 diabetes.
499 *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7581–7590 (2019).
- 500 42. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic Associations Created by Rare
501 Variants Do Not Explain Most GWAS Results. *PLoS Biol.* **9**, e1000579 (2011).
- 502 43. Petukhova, L. et al. Genome-wide association study in alopecia areata implicates both innate
503 and adaptive immunity. *Nature* **466**, 113–7 (2010).
- 504 44. Hakonarson, H. et al. A novel susceptibility locus for type 1 diabetes on Chr12q13 identified
505 by a genome-wide association study. *Diabetes* **57**, 1143–6 (2008).
- 506 45. Plagnol, V. et al. Genome-wide association analysis of autoantibody positivity in type 1
507 diabetes cases. *PLoS Genet.* **7**, e1002216 (2011).
- 508 46. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
- 509 47. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. Genome-
510 wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2
511 diabetes susceptibility. *Nat. Genet.* **46**, 234–44 (2014).
- 512 48. Kang, R. et al. EnhancerDB: a resource of transcriptional regulation in the context of
513 enhancers. *Database (Oxford)*. **2019**, (2019).
- 514 49. de Rie, D. et al. An integrated expression atlas of miRNAs and their promoters in human and

- mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).
50. L.J., C. et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking* (2015).
51. Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* (80-.). **348**, 648–660 (2015).
52. Duong, D. et al. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics* **33**, i67–i74 (2017).
53. van Arensbergen, J. et al. High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* **51**, 1160–1169 (2019).
54. Li, M. J., Yan, B., Sham, P. C. & Wang, J. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Brief. Bioinform.* **16**, 393–412 (2015).
55. Friedel, M., Nikolajewa, S., Sühnel, J. & Wilhelm, T. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.* **37**, D37–40 (2009).
56. Bishop, E. P. et al. A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem. Biol.* **6**, 1314–20 (2011).
57. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
58. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
59. Kalita, C. A. et al. High-throughput characterization of genetic effects on DNA–protein binding and gene transcription. *Genome Res.* **28**, 1701–1708 (2018).
60. Wang, X. et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *Elife* **5**, e10557 (2016).
61. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).

Acknowledgments

The authors 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 Research and Development Program (2016YFC0901603), the China 863 Program (2015AA020108), 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 G.G. was supported in part by the National Program for Support of Top-notch Young Professionals.

Part of the analysis was 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.

Authors' Contributions

G.G. conceived the study and supervised the research; F.Y.S. and Y.W. designed and implemented the computational framework as well as the website; F.Y.S., Y.W. and Y.L. completed the data curation; H.D., F.Y.S, N.L., and X.W.C. contributed to the independent luciferase reporter validation; F.Y.S. and G.G. wrote the manuscript with comments and input from all coauthors.

Competing Interests

The authors declare no competing interests.

Ethics approval and consent to participate

Not applicable

Consent for publication

569 Not applicable

570

571 **Supplementary information**

572 Supplementary Tables

573 Supplementary Figures

574

575

576 Figure 1

578 **Figure 2**

579

580

581 **Figure 3**

582

583

Figure legends

Figure 1: **a**, Overview of the CARMEN workflow. **b**, The ROC curve for the CARMEN-E module. The performance was evaluated on the testing dataset of the curated MPRA dataset with 738 variants (207 positive variants, 531 control variants). The ExPecto predicted scores were calculated as the average absolute score of 208 tissues. **c**, The ROC curve for the CARMEN-F module. The performance was evaluated on a pathogenic dataset with 2,135 variants (Supplementary Table 7). **d**, The performance comparison of CARMEN to other tools²⁶ bubbles colored by F1 score. The thresholds of the different tools were obtained from respective official websites or reference papers (Supplementary Table 10). **e**, The recall of different tools on the known expression-modulating variants²⁶. The X axis indicates the proportion of variants from the significant expression-modulating variants that were predicted as positive variants by the different tools. **f**, Performance comparison on a curated list of experimentally characterized variants validated by various low-throughput technologies from the literature (Supplementary Table 3). **g**, Performance comparison on an independent luciferase validated dataset⁶⁰ (Supplementary Table 3). 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 rs11119843⁶¹ is missed by all other tools.

Figure 2: CARMEN helps pinpoint the functional mechanism of noncoding expression-modulating variant. **a**, CARMEN predicted rs883868 as expression-modulating variant which was further validated by independent CRISPR/Cas experiment⁴¹. **b**, The CARMEN annotation indicated that the alternative allele C decreases the binding affinity of YY1, which was also validated via 3C-qPCR⁴¹. **c and d**, Mutation maps for the variant effects on YY1 transcription factor binding in GM12878 (**c**) and K562 (**d**).

Figure 3: Prioritizing causal variants over linkage variants with lead SNPs in GWAS. **a and c**, 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 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 best CARMEN score. **b and d**, Luciferase report assays well validated the prediction in type 1 diabetes mellitus (**b**) and type 2 diabetes (**d**). The results were from more than 8 technical replicates with 3 independent experimental replications. **e**, Comparing the performances of CARMEN and other tools on the two luciferase assays that validated as the positive variants above.

619 **Supplementary Figures**

620

621 **Supplementary Figure 1.** The variant annotations of the CARMEN framework.

622 Each input variant had 2,424 features, the genomic features were all annotated with sequence-based
 623 models, and the 8 evolutionary features included PhastCons and PhyloP scores, which were
 624 downloaded from the UCSC genome browser.

625

Supplementary Figure 2. Comparison of the training profiles between the CARMEN annotation module and DeepSEA.

(a) The training profiles of the convolutional neural network were the features with 2,403 chromatin profiles from ENCODE, which was greater than the number of training profiles used for DeepSEA (919). Moreover, DeepSEA did not incorporate the DNA methylation profiles. (b) The AUROC of the model performance of the chromatin profile features in CARMEN with 2,403 features; each feature has a separate prediction model, and the performance was evaluated on 15% of training data. The Y axis is the AUROC score.

Supplementary Figure 3. Classification performance of the common features of the CARMEN annotation module and DeepSEA via AUROC (a) and AUPRC (b).

There are 466 common features between the CARMEN annotation module and DeepSEA; the performance of the CARMEN common feature prediction models were evaluated on independent testing sets, and the performance of DeepSEA is extracted from the original paper¹².

Supplementary Figure 4. Illustration of data merging, converting to FASTA format and one-hot encoding.

Two sequences from two data files overlap each other, and the second sequence has a higher peak value; as a result, it is included in the dataset. Files in bed format are converted to FASTA format.

An illustration of one-hot encoding: each sequence is represented by a 200×4 binary matrix, with columns corresponding to A, C, G and T.

Supplementary Figure 5. Pipeline for predicting variant effects with sequence-oriented models. 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, a number between 0-1 is generated, which represents the predicted effects in term of log fold change between Ref and Alt sequences.

Supplementary Figure 6. Feature selection procedure.

The variants with 2,424 annotated features were used as the input to pretrain a multilayer neural network, then DeepLIFT was used to estimate the contribution scores and obtain the important features for discriminating functional variants.

Supplementary Figure 7. CARMEN performance on a cancer-risk dataset and the BiT-STARR-seq dataset.

Comparing CARMEN with other tools to prospect the noncoding-regulation modulating variants. (a) The receiver operating characteristic (ROC) curves for the cancer-risk dataset. Since ExPecto provides the scores of the 208 tissues for each variant, we did not calculate the AUROC and AUPRC of ExPecto. (b) The ROC curves for the BiT-STARR-seq dataset. After curating 43,500 SNPs from Kalita, Cynthia A., et al.⁵⁹ as the second dataset, CARMEN still had the best performance. Among 43,500 variants, 2,720 SNPs with an FDR less than 10% were considered significant ASE variants, and this dataset was called the BiT-STARR-2018 dataset. CARMEN had the highest AUROC (0.646) on this dataset. (c) CARMEN performance using the best AUROC in the BiT-STARR-seq dataset as the threshold. Since this independent testing dataset is imbalanced with a 1:15 ratio of positive to control variants, we calculated the weighted accuracy, which balanced the sensitivity and specificity, and the F1 score, which balanced precision and recall, to compare CARMEN performance to that of the other tools. Although CARMEN had a comparable

679 F1 score with that of the GWAVA Unmatched model, CARMEN had the highest weighted
 680 accuracy (0.579) among all tools tested; it performed particularly well in terms of sensitivity. The X
 681 axis indicates the weighted accuracy, the Y axis represents different tools, and the bubble color
 682 represents the F1 score. The thresholds of different tools were supplied by the official tools' website
 683 or reference papers (ordered by weighted accuracy). (d) CARMEN performance, using the best
 684 AUROC in the cancer risk dataset as the threshold.
 685

Supplementary Figure 8. The distribution of GWAS lead SNPs with low CARMEN scores

The lead SNPs in the GWAS Catalog, which have CARMEN scores lower than 0.005 in the European, Han-Chinese and African populations, were selected. Then, we calculated the CARMEN score difference between the selected lead SNPs and their high linkage SNPs ($r^2 > 0.75$). 45.33% of the selected variants had lower CARMEN scores than LD SNPs with the threshold of 0.005. While most of the differences are modest, 6.65% of variants show differences larger than thirty-fold.

Supplementary Figure 9. Feature importance proportion.

Feature importance was calculated separately in the two modules. The value of the different color bar represents the cumulative importance of each feature category.

Supplementary Figure 10. UCSC Genome Browser annotations for variants.

(a) Annotations for rs1701704. This variant falls into the intron of gene IKZF4 with low H3K27ac modification but without annotation about DNase cluster and transcription factor binding. (b) Annotations for rs705698. (c) Annotations for rs1727313. (d) Annotations for rs146239222.

Supplementary Figure 11. Motifs matched for the sequence around rs705698.

Flanking sequences of 15 bp of SNP rs705698 were used to calculate the mutation score. Features annotated by CARMEN with absolute log fold change larger than 1 were selected. The red block indicates that this mutation will increase the binding of this feature. The blue block indicates that this mutation will decrease the binding of this feature. Mutation scores were scaled in each feature. The top heatmap is the mutation score of sequence with reference allele and the bottom heatmap is the mutation score of sequence with alternative allele.

Supplementary Figure 13. The time cost of CARMEN.

Estimating the running time of CARMEN with different number of variants. With 5 times random sampling of different numbers subset from 43,500 variants. Error bar represents mean $\pm$ sd.