Published online 10 October 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 22* ***e146***
doi: 10.1093/nar/gkz868

Convolutional neural network model to predict causal risk factors that share complex regulatory features

Taeyeop Lee[1,†], Min Kyung Sung[2,3,†], Seulkee Lee[1,2,4,†], Woojin Yang[2,5], Jaeho Oh[2], Jeong Yeon Kim[2], Seongwon Hwang[6], Hyo-Jeong Ban[7] and Jung Kyoon Choi[2,*]

[1]Graduate School of Medical Science and Engineering, KAIST, Daejeon 34141, Republic of Korea, [2]Department of Bio and Brain Engineering, KAIST, Daejeon 34141, Republic of Korea, [3]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK, [4]Department of Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, Republic of Korea, [5]Korean Bioinformation Center (KOBIC), KRIBB, Daejeon 34141, Republic of Korea, [6]Seminar for Statistics, Eidgenössische Technische Hochschule (ETH) Zurich, CH-8092 Zurich, Switzerland and [7]Future Medicine Division, Korea Institute of Oriental Medicine, Daejeon 34054, Republic of Korea

Received August 30, 2019; Editorial Decision September 23, 2019; Accepted October 01, 2019

ABSTRACT

Major progress in disease genetics has been made through genome-wide association studies (GWASs). One of the key tasks for post-GWAS analyses is to identify causal noncoding variants with regulatory function. Here, on the basis of >2000 functional features, we developed a convolutional neural network framework for combinatorial, nonlinear modeling of complex patterns shared by risk variants scattered among multiple associated loci. When applied for major psychiatric disorders and autoimmune diseases, neural and immune features, respectively, exhibited high explanatory power while reflecting the pathophysiology of the relevant disease. The predicted causal variants were concentrated in active regulatory regions of relevant cell types and tended to be in physical contact with transcription factors while residing in evolutionarily conserved regions and resulting in expression changes of genes related to the given disease. We demonstrate some examples of novel candidate causal variants and associated genes. Our method is expected to contribute to the identification and functional interpretation of potential causal noncoding variants in post-GWAS analyses.

INTRODUCTION

During the last decade, numerous efforts have been made to elucidate the genetic mechanisms underlying complex disorders. Major progress was made through genome-wide association studies (GWASs). However, developing methods to pinpoint the DNA variants that actually increase the risk of the associated disease is a major challenge that GWASs still face (1). GWASs cannot pinpoint causal disease variants but can only report linkage disequilibrium (LD) blocks including many neutral SNPs linked to causal loci. To exacerbate the problem, the majority of disease-associated DNA variations are thought to alter not the gene itself but the regulation of gene expression (2). Our incomplete knowledge of noncoding regions limits the functional interpretation of underlying DNA variants. Fortunately, the wealth of cell-type-specific human epigenomes help with the identification of functional noncoding variants (1,3–5).

Deep learning is a powerful approach for learning complex patterns (6) and has been applied in genomics especially for deciphering the complexity of noncoding DNA sequences. For example, a deep-learning model that predicts regulatory codes for RNA splicing has been developed (7). As one of the most successful deep learning architectures, convolutional neural networks (CNNs) have been used to systematically learn the sequence motifs or regulatory patterns embedded in genomic regions recognized by DNA- or RNA-binding proteins (8) or in DNase I hypersensitive sites (DHSs) (9).

In this study, we developed a deep learning framework based on CNNs to discover regulatory variants that may play a causative role in increasing the risk of the five major psychiatric disorders and four autoimmune diseases: autism spectrum disorder (ASD), attention deficit-hyperactivity disorder (ADHD), bipolar disorder (BPD), major depressive disorder (MDD), schizophrenia (SCZ), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), Crohn's disease (CD) and ulcerative colitis (UC). We utilized numerous functional features while combining them nonlinearly to model complex patterns shared by risk

[*]Correspondence should be addressed. Tel: +82 42 350 4327; Fax: +82 42 350 5367; Email: jungkyoon@gmail.com
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

variants. We were able to discover novel candidate SNPs that may actually contribute to disease development.

MATERIALS AND METHODS

Identification of association blocks

In order to make the input data for the CNN model, we first identified chromosomal association blocks. To this end, we employed a GWAS dataset for five major psychiatric disorders and four autoimmune diseases (Supplementary Table S1) (10,11). The association P values were downloaded from the Psychiatric Genomic Consortium portal (https://www.med.unc.edu/pgc/results-and-downloads) for psychiatric disorders. As for the autoimmune diseases, the association P values were retrieved from the largest meta-analysis for each disease (12–14). For imputation, we used the EUR (European ancestry) samples of the 1000 Genomes Project phase 1 release (version 3) as the reference panel (15). Imputation of summary statistics was performed by using ImpG-Summary (16). The SNPs that showed the strongest associations and were at least 1 Mb apart from one another were defined as lead SNPs. To identify the lead SNPs, we sorted all SNPs across the whole genome according to the observed or imputed P values and picked SNPs from the top of the list while maintaining a >1 Mb distance from each of the previously selected ones. We then searched upstream and downstream regions flanking each lead SNP for the 30 most significant SNPs. We then discarded those with $P > 5 \times 10^{-4}$. In this manner, we identified association blocks carrying the lead SNP and their neighboring SNPs while constraining the maximum number of neighboring SNPs to 30 (Figure 1A). The resulting number of association blocks was 340, 391, 474, 405 and 601 in ADHD, ASD, BPD, MDD and SCZ for psychiatric disorders, and 435, 849, 431 and 383 in RA, SLE, CD and UC for autoimmune diseases, respectively (Supplementary Table S2).

Feature set construction

The overall processes for our feature map construction are illustrated in Supplementary Figure S1. We obtained open chromatin profiles in 349 different samples as provided in the form of DHS peaks by the ENCODE Project (17) and the Roadmap Epigenomics Project (18). A total of 606 histone modification profiles from the Roadmap Epigenomics Project (18) were obtained as the narrow peak bed files in 124 different biological conditions. The KEGG database (19) was utilized to incorporate gene functions as features. Each SNP was mapped to their target gene and its KEGG pathway when they resided in the gene body or 500 kb upstream of the TSS. We ran FIMO (http://meme-suite.org/doc/fimo.html) (20) to search the TRANSFAC (21) and JASPAR (22) databases of position weight matrices for 996 transcription factors (TFs). We used the P value threshold of 10^{-4} for feature mapping. As a result, we compiled 2,252 functional features consisting of DHSs, histone modifications, KEGG pathways, and TF binding sites. Because using too many features results in overfitting, we filtered features that were shared by only a small number of SNPs in the association blocks (Supplementary Figure S1). Specifically, we retained the features that were mapped to any SNPs in > 95% of the association blocks in each disease model. The resulting number of features was 714 for ADHD, 711 for ASD, 730 for BPD, 714 for MDD, 739 for SCZ, 791 for RA, 741 for SLE, 845 for CD and 834 for UC.

Model design

The developed CNN model consists of hierarchical pattern detectors that learn the regulatory features commonly present in disease-associated genomic loci. In our model, we used two convolution layers; (i) the first layer acts as a local feature extractor at the individual SNP level and (ii) the second layer combines the detected patterns into more complex biological features.

Convolution layer 1. The convolution layer 1 takes an input matrix (X_{mn}) with a size of $M \times N$as described below:

$$X_{mn} = \begin{cases} 1 & \text{if the } m\text{-th features is associated with the } n\text{-th SNP} \\ & \quad (\text{for all } 1 \le m \le M \text{and } 1 \le n \le N) \\ 0 & \text{otherwise} \end{cases} .$$

M is the number of regulatory features that survived the filtering step described above, and N is the total number of candidate SNPs in an association block. With this input matrix and tunable patterns represented as kernel w^k, the first feature map array (${}^1c_n^k$) was obtained by the convolution process,

$$ {}^1c_n^k = \sum_{m=1}^{M} w_m^k X_{mn}, $$

where k is the index for 50 filters used in our model ($1 \le k \le K, K = 50$). Each convolution kernel w^k is a weight vector with a length of M. This process is equivalent to performing a one-dimensional convolution (8) with a moving window of step size 1 on the list of consecutive N SNPs each having M channels. As implied in the above formula, K types of pattern detectors were used for each SNP without considering the effect of neighboring SNPs. After convolving the matrix X_{mn}, we added a bias vector, b_k, which was expected to make the model more flexible and allow the model to solve given problems more effectively. Subsequently, a rectified linear unit (ReLU) was applied as follows:

$$ h_n^k = \mathrm{ReLU}({}^1c_n^k + b_k), $$

$$ \text{where } \mathrm{ReLU}(x) = \begin{cases} x & \text{if } x \ge 0 \\ 0 & \text{if } x < 0 \end{cases} . $$

We then stacked the output vector h_n^k to compose a matrix H_{kn} with a size of $K \times N$, which corresponds to scores measuring how well the features of each SNP match the patterns of the shared weights.

Convolution layer 2. The convolution layer 2 operates on the output matrix of the previous layer (H_{kn}). The second feature map vector (${}^2c_n$) was obtained as

$$ {}^2c_n = \sum_{k=1}^{K} w'_k H_{kn}. $$

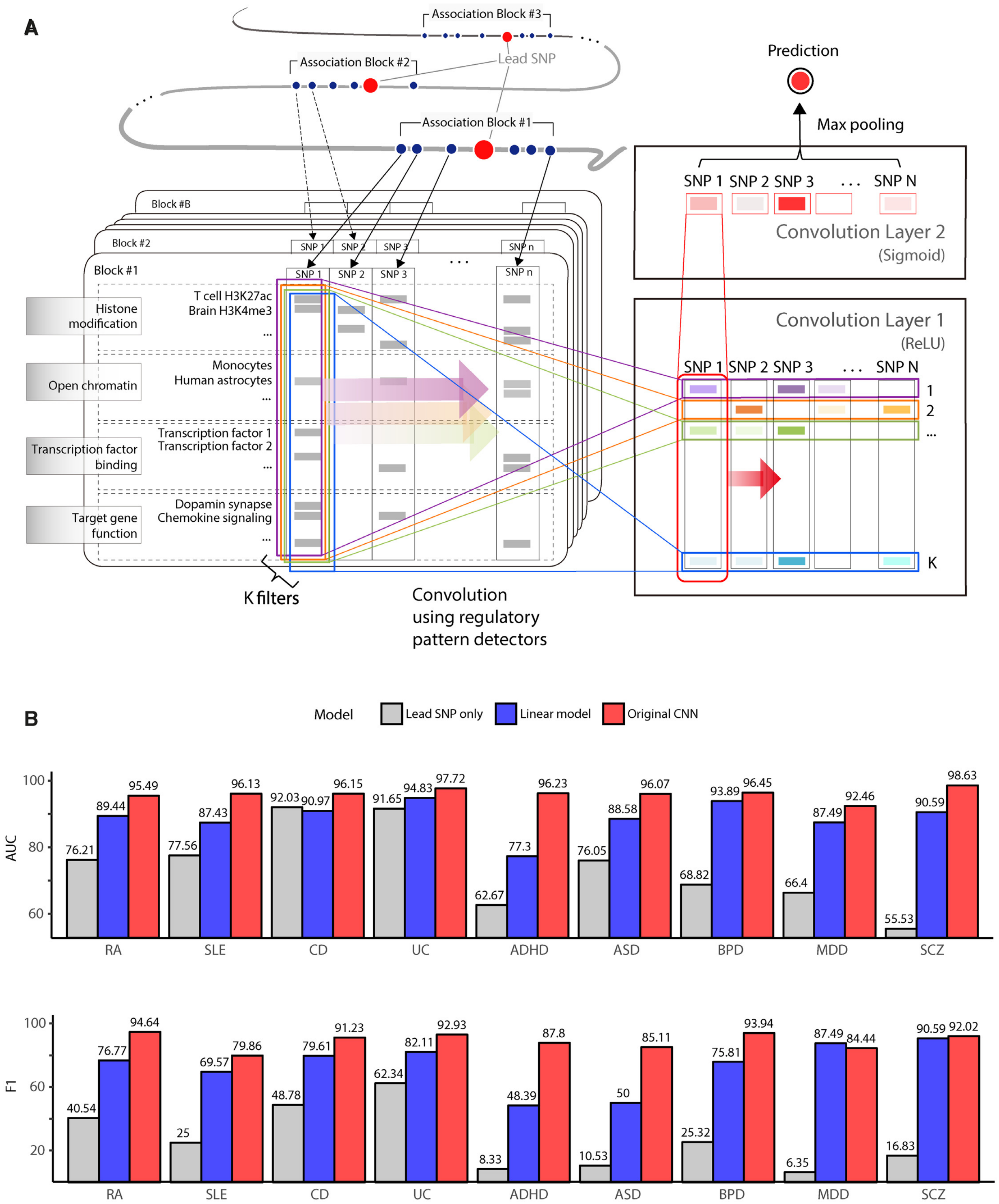

Figure 1. Model schematic and performance. (**A**) CNN framework to detect regulatory patterns shared by risk variants residing in multiple association blocks centered on lead SNPs. In this example, block #1 carries *n* SNPs including the lead SNP. We apply *k* different kernels that learn particular patterns composed of various regulatory features encompassing DHSs, histone modifications, target gene function, and TF binding sites. At this stage, an autoencoder is used for pre-training. In this manner, the first convolution layer scores *n* SNPs with *k* pattern detectors. Afterward, another convolution layer is applied to combine the *k* scores, thereby enabling nonlinear combinatorial modeling of regulatory patterns. The output of the second layer serves as the prediction score for each SNP. The model is trained to maximize the likelihood derived from the block scores that are assigned by max pooling. (**B**) Model performance of rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), Crohn's disease (CD), ulcerative colitis (UC), attention deficit-hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BPD), major depressive disorder (MDD) and schizophrenia (SCZ) measured on the basis of AUC and F1. The red, blue and gray bars are for the original CNN model, linear model with only one convolution layer, and model with only the lead SNPs. Model training and performance evaluation were carried out on the training, validation, and testing sets (Supplementary Figure S2 and Supplementary Table S2).

In this step, only one tunable weight vector was used to linearly combine the K patterns for high-level feature scoring of each SNP. Then, we added a bias term b' and scaled ${}^{2}c_n + b'$to the 0–1 range by the sigmoid function:

$$o_n = \text{sigmoid}\,({}^{2}c_n + b'),$$

$$\text{where sigmoid}\,(x) = \frac{1}{1 + e^{-x}}.$$

o_n can be regarded as the prediction score of the n-th SNP. Prediction scores close to 1 indicate that certain common regulatory patterns are embedded in the features of the given SNPs. Finally, we applied max-pooling by taking the maximum of o_n as

$$\text{pooling}(o_n) = \max(\{o_1, o_2, ..., o_N\}).$$

This corresponds to the per-block score of the SNP whose features best match the common patterns shared by different association blocks.

Model training

The input data for the model was split into training, validation, and testing sets (Supplementary Figure S2 and Supplementary Table S2). The validation and testing sets were used to select the best hyperparameter set and report final performance levels, respectively. We trained the model parameters to minimize negative log likelihood (NLL) defined as follows:

$$\text{NLL}(\theta) = -\frac{1}{B}\sum_{s=1}^{B}(Y^s \log f(\theta)^s + (1 - Y^s)\log(1 - f(\theta)^s))$$

where $f(\theta)^s$ is an output for the sth block in a mini-batch of size B from the training set ($B = 100$ was used for all experiments). Each target, Y, serves as the dependent variable and can be either 1 (for true cases or the blocks that carry the pattern of SNPs associated with the disease) or 0 (for false cases or the blocks that are unlikely to carry the pattern of SNPs associated with the disease). For each SNP, false cases were constructed by shuffling the features. In doing so, it is anticipated that the features of false cases do not reflect any disease-related common patterns. The shuffling procedure was done separately for each feature group. We generated false cases that are ten times the true cases. The NLL was composed of parameters (θ) including w_m^k, b_k, w'_k and b', which were updated by the standard back-propagation algorithm with momentum. To prevent overfitting, early stopping and pre-training by an autoencoder were used. More details of the model training processes are documented in Supplementary Information.

Feature importance analysis

We employed random forest to assess the relative importance of each feature (23). The input SNPs for training RF were labeled as positive (prediction score > 0.5) or negative (prediction score < 0.5) according to our CNN results. For each feature, we calculated the Mean Decrease Gini (MDG), which is defined as the average of total decrease in Gini impurities in each tree. A greater MDG indicates higher importance of a feature. The empirical P value of MDG was calculated by 1000 random permutations. The hypergeometric distribution was used to test whether disease-related features stood out in terms of the MDG. Neural and immune features were defined on the basis of open chromatin or histone marks in normal neuronal cells or tissues and in immune-related blood tissues (all lymphoid cells and granulocytes) or cell lines, respecitvely. KEGG pathways including ‘nervous system’, ‘neurodegenerative disease’, ‘substance dependence’, ‘inflammatory process’, ‘immune process’, and ‘chemokine signaling pathway’ were also included. As negative controls, we used irrelevant features consisting of open chromatin peaks, histone modifications, and KEGG pathways related to the digestive and circulatory system in the same way. In our network analysis, the neural features with a significant ($P < 0.05$) MDG were selected for visualization. RF model building and feature importance analysis were implemented by using R packages randomForest and rfPermute (24). Network visualization of the significant neural features was based on Cytoscape (v3.5.1) (25).

Functional analyses of predicted variants

Details on the TF binding and allelic imbalance analysis, evolutionary conservation analysis and target gene function analysis are provided in Supplementary Information. Additional dataset not used in the training process was employed for the analysis of the autoimmune diseases. A total of 100 histone modification profiles in 16 blood tissues or cell lines were obtained as peak bed files from the BLUEPRINT epigenome project (www.blueprint-epigenome.eu) (26).

RESULTS

Prediction model and performance

Our CNN model was trained on the feature vectors across multiple association blocks (Figure 1A and Supplementary Figure S1). The dependent variable of the model is 1 for true cases or the blocks that carry the pattern of SNPs associated with the disease and 0 for false cases or the blocks that do not carry the pattern of SNPs associated with the disease. The premise of the model is that there are one or more functional variants in association blocks, and that many of the variants share certain patterns of regulatory features despite being scattered in different blocks. Therefore, the association blocks identified through GWASs served as true cases.

In analogy, a true case (association block) can be compared to a face image, and SNPs can be compared to eyes, nose, or mouth. By observing many face images, a CNN model can learn that a face has eyes, nose, and mouth in common, and decide whether a given picture is a face or not. Likewise, if a CNN model is trained with multiple true cases, it can learn that association blocks carry SNPs with certain patterns and can decide whether a given region is a true case.

The number of association blocks used as true cases for ADHD, ASD, BPD, MDD, SCZ, RA, SLE, CD and UC was 340, 391, 474, 405, 601, 435, 849, 431 and 383, respectively (Supplementary Table S2). These blocks were partitioned into the training set, validation set, and testing set (Supplementary Table S2). Details of the learning processes are summarized in Supplementary Figure S2. Performance evaluation was based on the area under the receiver-operator characteristic curve (AUC) and F1 value (Figure 1B). We modified our model to learn the features of only the lead SNPs (i.e. the most significant SNPs indexing each association block) or to learn patterns composed of the linear combinations of features (by using only one convolution layer). The lowered performance of the modified models (gray and blue bars of Figure 1B) indicates that common regulatory patterns need to be searched for through all variants in each chromosomal block in a complex, non-linear fashion. This justified the usage of a CNN model for this task.

Overall, the lowest performance was achieved for MDD, probably reflecting that genetic factors play a less significant role relative to the other diseases (27). We defined positive calls as variants that were assigned a prediction score greater than 0.5. The list of these putatively causal variants in each disease is provided in Supplementary Table S3.

Biological validation of prediction results

First, true causal variants are expected to have a certain level of statistical association with the given phenotype. Indeed, our model assigned higher prediction scores to associated variants in the testing set, which is independent of the training processes (Supplementary Figure S3). In >50% of the chromosomal blocks with at least one positive call, the variant with the strongest statistical association (i.e. lead SNP) was positively predicted (Supplementary Table S2). Also, there were many cases in which the greatest prediction score was assigned to the lead SNPs (Supplementary Figure S4). Of importance, our prediction method was able to single out one of statistically indistinguishable variants (compare red diamonds and blue circles in Figure 2).

Second, disease-related features are expected to play an instrumental role when predicting causal variants. For example, psychiatric disorders should be associated with brain-related features while autoimmune diseases with immune-related features. To test this, we employed the random forest classifier to assess the contribution of each feature to the prediction processes. By randomizing each feature, the explanatory power of the given feature in discriminating positive and negative calls could be estimated. We used the MDG score for this measure (see Materials and Methods). With this metric, we observed higher discriminative power for neural features and immune features than irrelevant features in the psychiatric disorders and autoimmune diseases, respectively (Figure 3A).

In addition, some neural features reflected the pathophysiology of the relevant psychiatric disorder (Figure 3B). For example, astrocyte (red nodes) and dorsolateral prefrontal cortex (green nodes) are often implicated in ASD (and ADHD) and SCZ (and BPD), respectively. Fetal features (blue nodes) were important when characterizing neurodevelopmental disorders such as ASD and SCZ. Similarly, the BLUEPRINT epigenome data for various immune cell types were used for the analysis of the autoimmune disease results. Positive calls were more enriched in the regulatory regions of lymphocyte lineages rather than granulocytes. This is in good agreement with the pathophysiology of autoimmune diseases (Figure 3C and Supplementary Figure S5). Moreover, a comprehensive target gene analysis also supported the clinical relevance of our prediction (Supplementary Figure S6).

We assessed the importance of features also by simply examining the relative weight ranking of biological features in each kernel (see Supplementary Information). Neural and immune features in psychiatric disorder and autoimmune disease, respectively, showed significantly higher level of importance compared to other features, in a subset of kernels (Supplementary Figure S7).

Third, causal noncoding variants are likely to lie in the regulatory regions of relevant tissue types. We first tested whether positive calls are enriched in regulatory regions derived from independent data. The BLUEPRINT epigenome data for various immune cell types were useful for this purpose because they were not used for model training. The fractions of positive calls for the autoimmune diseases were significantly higher than negative calls in H3K4me1 and H3K27ac regions from the BLUEPRINT epigenome data (Figure 4A). Next, to compare disease-relevant tissues with others, we ordered all available tissue types from the Epigenome Roadmap project depending on the degree of positive call enrichment. Expectedly, positive calls were enriched in the disease-related tissues for both the psychiatric disorders and autoimmune diseases (Figure 4B).

Fourth, the mechanisms by which noncoding variants contribute to disease phenotypes should involve TF binding changes. For this test, we utilized TF footprint data generated by base-resolution DHS analyses (28). The positive hits included a significantly higher fraction of nucleotides that are in contact with cognate TFs than the negative calls (Figure 4C). A useful method to test the functionality of regulatory variants is to examine allelic imbalance in chromatin accessibility (29). By examining allelic patterns in footprint reads, we found that different alleles at positive calls are more likely to cause distinct regulatory variation (Supplementary Figure S8A). We also tested whether the predicted putative causal variants tend to affect gene expression levels. When examined using the 1000 Genomes whole-genome and transcriptome data, positive calls showed higher levels of expression association, further supporting the functionality of the predicted variants at the transcription level (Supplementary Figure S8B).

Finally, true causal variants for major psychiatric disorders are likely to reside in regions that are critical for brain development and function. Therefore, one can anticipate higher evolutionary conservation, especially among primates, for regions surrounding the putative causal variants. Indeed, the odds of positive calls in conserved sequences were statistically significant (Figure 4D). This tendency was more distinct with ASD, ADHD and SCZ, in which aberrations in neural development play a critical role, as compared to MDD and BPD. Also, the average degree of sequence conservation was markedly higher for genomic regions centered on positive calls than negative calls (Supplementary Figure S9A). The degree of conservation was less

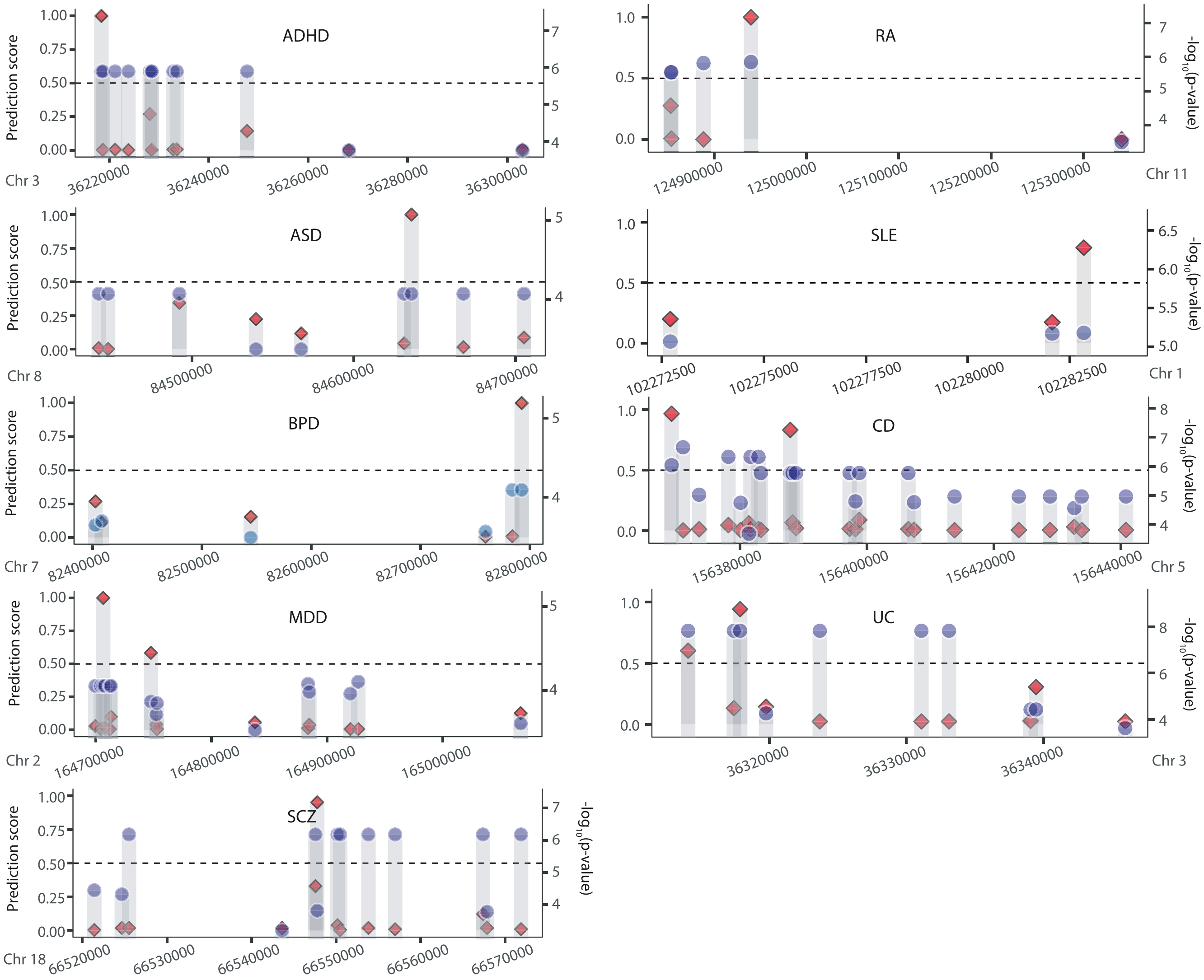

Figure 2. Comparison of our prediction scores (red diamonds on the left y-axis) and association statistics (blue circles on the right y-axis) for individual SNPs in exemplary association blocks. Our functional prediction enables to discern statistically indistinguishable variants.

significant when measured among mammals or vertebrates (Supplementary Figure S9B).

Additionally, we sought to test how other existing tools perform in the above biological validations in comparison to our CNN results. Among the tools that predict the pathogenicity of variants, we used CADD (30), PhastCons (31), and LINSIGHT (32). In contrast to the CNN model, pathogenic variants defined by CADD, PhastCons, and LINSIGHT showed weaker or unclear enrichment patterns in disease-related cell types from the Epigenome Roadmap project for both psychiatric disorders and autoimmune diseases (Supplementary Figure S10, compare with Figure 4B). We also assessed overlap with TF footprints (Supplementary Figure S11, compare with Figure 4C), allelic imbalance in chromatin accessibility (Supplementary Figure S12, compare with Supplementary Figure S8A), and association with gene expression levels (Supplementary Figure S13, compare with Supplementary Figure S8B). According to these results, we conclude that the positive calls predicted by the CNN model are biologically more meaningful than the pathogenicity calls by the other tools.

Examples of novel candidate causal variants

As shown in Figure 1B, our method was able to detect common patterns shared by variants other than the lead SNPs of association blocks. In other words, there must be numerous cases in which the tag SNPs or lead SNPs detected by the typical GWAS analysis based on statistical association may not act as causal variants for the disease. For example, a known GWAS variant of RA (33), rs773125, located on chromosome 12, has the strongest association with the RA phenotype (Figure 5A). Previous GWAS studies assigned this SNP to CDK2 on the basis of physical distance. However, the nearest gene is not always the actual target gene. Only a small fraction of distal enhancers target the nearest transcript (34). Not surprisingly, CDK2 has no clear role in association with RA. Furthermore, rs773125 is not

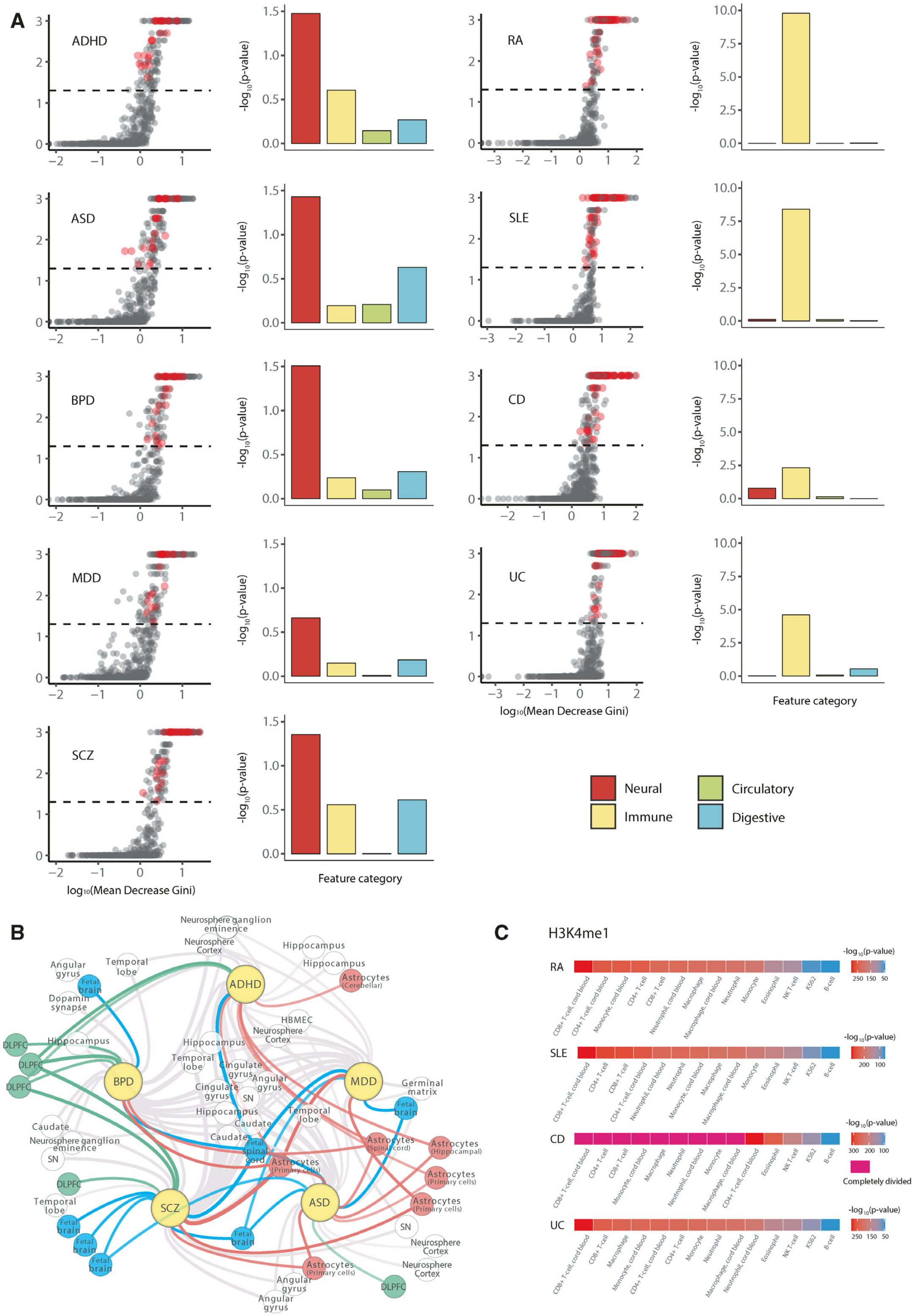

Figure 3. Pathophysiological relevance of prediction outcomes. (**A**) Results of feature importance analysis. The MDG score (x-axis) and its *P* value (y-axis) for individual features with significant ($P < 0.05$) disease related features highlighted in red (left panel). Enrichment of the significant features in different categories (neural, immune, circulatory, or digestive) as estimated by the hypergeometric test (right panel). (**B**) Network of neural features that were important in each disease model. The small nodes represent neural features that showed a significant ($P < 0.05$) MDG in the prediction model of the connected disease. The significant neural features, including those related to fetal brain (blue), astrocytes (red), and dorsolateral prefrontal cortex (green), were mapped to the relevant disorder (yellow). (**C**) Enrichment of positive calls for autoimmune diseases for the BLUEPRINT epigenome data for various immune cell types. Positive calls are more enriched in the regulatory regions of lymphocyte lineages rather than granulocytes. Shown here are regulatory regions marked by H3K4me1. The H3K27ac data is provided in Supplementary Figure S5.

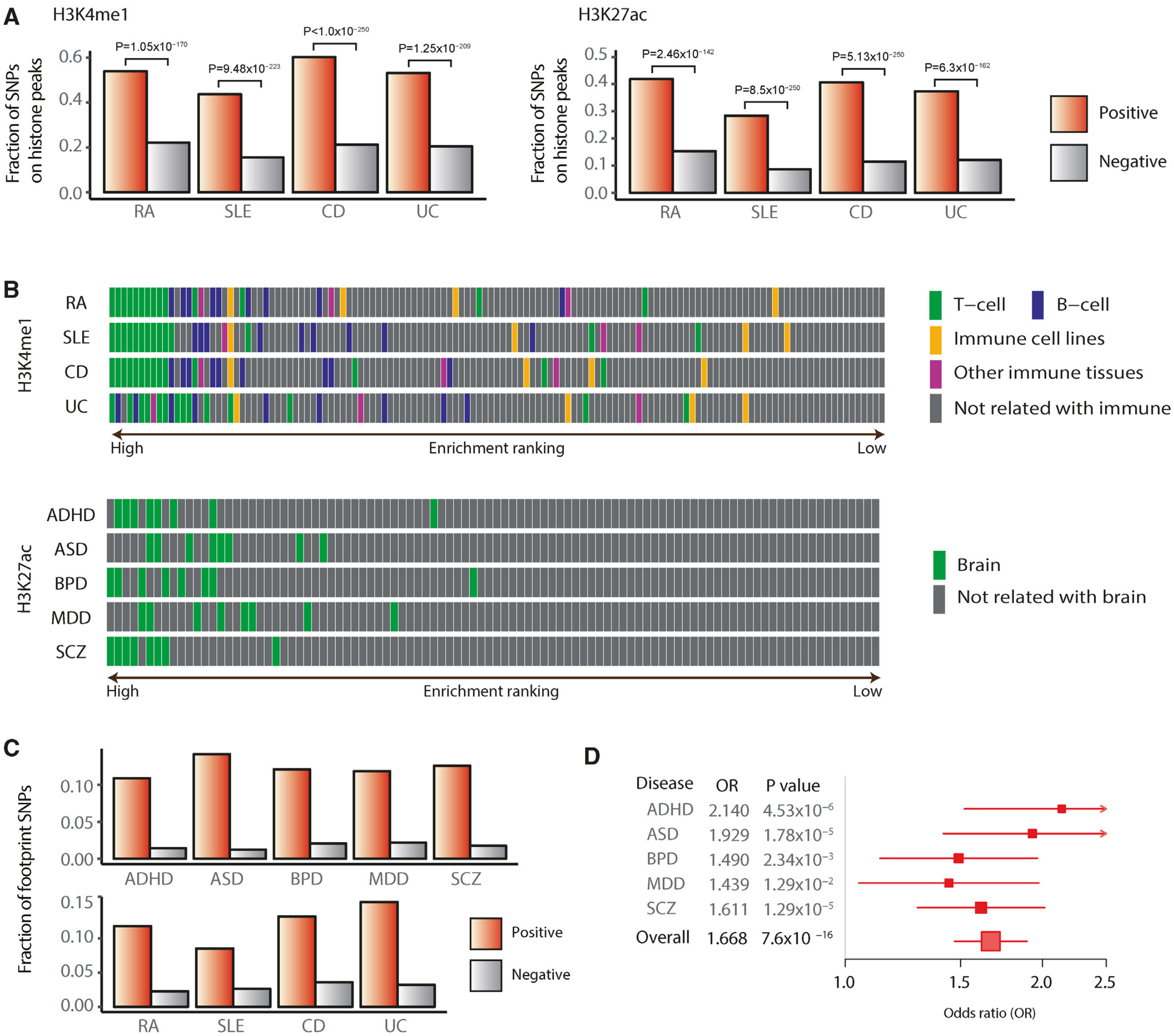

Figure 4. Functional relevance of prediction outcomes. (**A**) Fractions of positive calls and negative calls for autoimmune diseases in H3K4me1 and H3K27ac regions from the BLUEPRINT epigenome data. (**B**) Enrichment of positive calls for autoimmune diseases in immune-related cells and psychiatric disorders in brain-related cells from the Epigenome Roadmap project. (**C**) Proportion of positive (prediction score > 0.5) SNPs and negative (prediction score < 0.5) SNPs that match TF footprints. (**D**) The ratio of the odds of positive calls in conserved regions to their odds in non-conserved regions. For the conserved regions, we searched association blocks for the primate PhastCons score > 0.5. The odds of positive calls were computed as the ratio of the positive to negative SNPs in the conserved or non-conserved regions. Shown is the odds ratio together with its 95% confidence interval and *P* value.

located in active regulatory regions marked by H3K4me1 or H3K27ac in any types of immune-related tissues (Supplementary Figure S14A). Our prediction was negative on rs773125. Instead, there were two positive SNPs (rs773114 and rs1873914) that showed a lower association with RA. GM12878 capture Hi-C data (35) located these SNPs in an enhancer region of RPS26 (Figure 5A). This site was an active regulatory region of CD14+ monocyte as well. There were several studies that implicated RPS26 in autoimmune diseases as a possible factor that can evoke autoimmunity (36,37).

We were able to find similar examples in psychiatric disorders. For example, rs150721234 has the strongest association strength with the SCZ phenotype in the LD block located on chromosome 10 (Figure 5B). This SNP resides in an intron region of C10orf68, which has no clear role in association with SCZ. Moreover, rs150721234 is not located in active regulatory regions marked by H3K4me1 or H3K27ac in any types of brain-related tissues (Supplementary Figure S14B). Our prediction was negative on rs150721234. Instead, there was a positive SNP (rs117885390) that showed a lower association strength

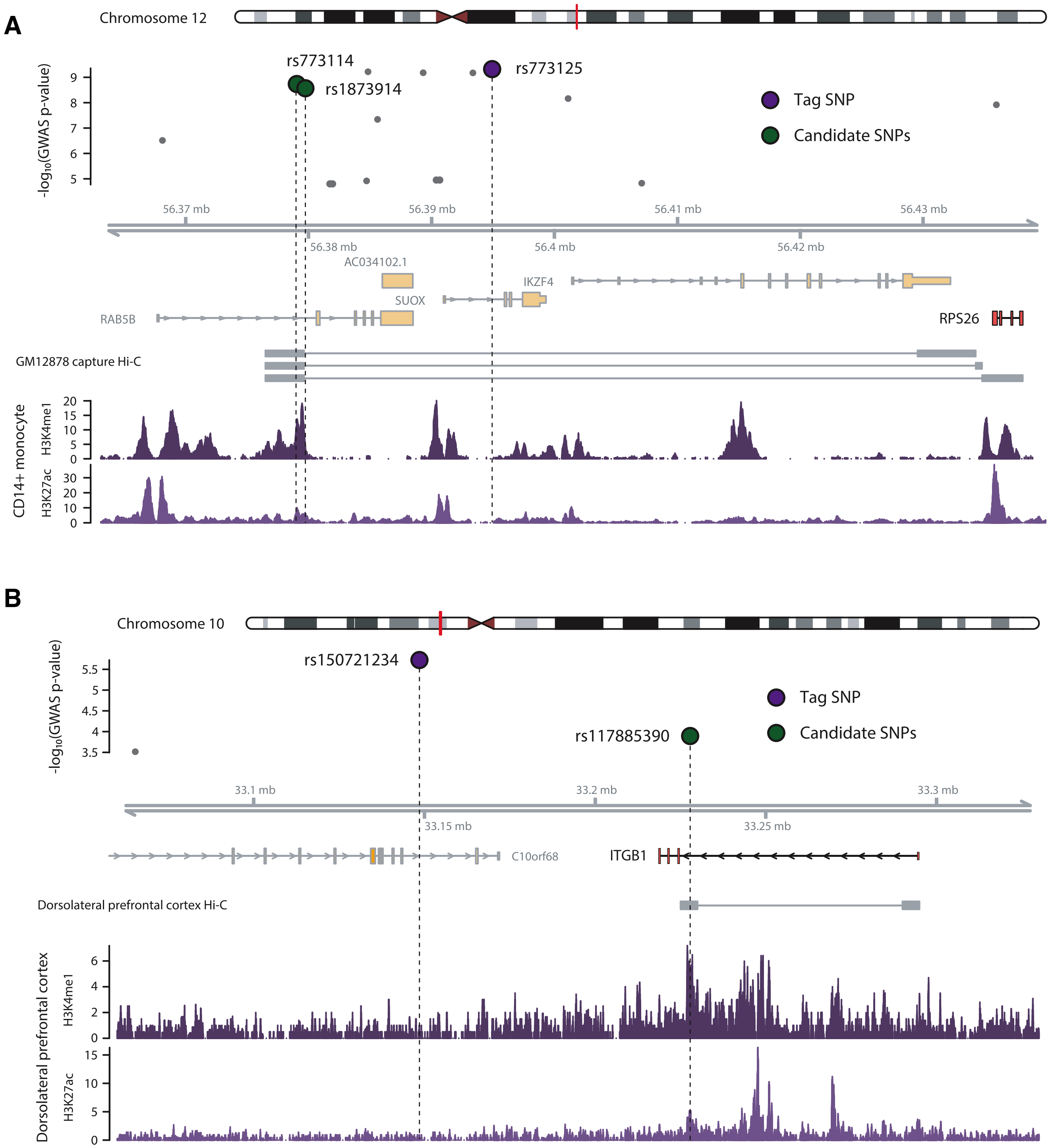

Figure 5. Examples of novel candidate causal SNPs. (**A**) Example for autoimmune diseases. Known GWAS SNP rs773125, located on chromosome 12, has the strongest association with RA. However, this SNP was a negative call from our prediction and not located in active regulatory regions in any immune-related tissues. The two positive SNPs (rs773114 and rs1873914) were located in an active regulatory region of CD14+ monocyte and connected to RPS26 according to GM12878 capture Hi-C data (35). (**B**) Example for psychiatric disorders. rs150721234 has the strongest association strength with SCZ in the LD block located on chromosome 10. However, this SNP was a negative call from our prediction and not located in active regulatory regions in any brain-related tissues. Instead, there was a positive SNP (rs117885390) that showed a lower association strength with SCZ. Dorsolateral prefrontal cortex Hi-C data (38) located this SNP in an enhancer region of ITGB1. This site was an active regulatory region of the dorsolateral prefrontal cortex tissue.

with SCZ. Dorsolateral prefrontal cortex Hi-C data (38) located this SNP in an enhancer region of ITGB1 (Figure 5B). This site was an active regulatory region of the dorsolateral prefrontal cortex. There were several studies showing that ITGB1 gene has a possible role during the development of schizophrenia (39, 40). These cases show that our method may be able to pinpoint functional variants that could be the actual cause of diseases among all variants associated with tag SNPs. Furthermore, these examples illustrate that once candidate variants other than tag SNPs are identified, one may be able to specify novel target genes that may shed light on the pathophysiology of the relevant phenotypes.

DISCUSSION

Our prediction model is different from the conventional architecture of CNNs intended for image processing. For biological reasons, we perform one-dimensional convolution while using a vector instead of a matrix for kernels. Only one-dimensional convolution is applicable for our purpose because genetic information is encoded in linear DNA strands. This type of convolution has been applied for predicting the sequence motifs of DNA- and RNA-binding proteins (8). While binding motifs are consecutive nucleotides that can be represented as a positional matrix, the order of SNPs along the chromosome does not carry meaningful biological information. This is why only vector kernels were applicable for our purpose. The power of our method stems from incorporating feature data that comes with external annotation. These features were not learned from DNA sequences *ab initio* but were incorporated on the basis of domain knowledge, which probably contributed to achieving high performance with a relatively small number of convolution layers. In addition, biological annotation helped with the validation and interpretation of prediction results. However, it must be noted that the model is based on the premise that at least one causal variant exists in the locus. Therefore, the presence of false positive GWAS signals may lead to undermine the performance of our approach.

Statistical approaches for fine-mapping are not applicable for rare variants because of limited power. In contrast to statistical fine-scale mapping, our prediction method is applicable to rare variants for which statistical association is difficult to estimate. This is important because the combined effects of rare variants may explain a significant proportion of genetic susceptibility to common diseases or traits (41–43). Expression quantitative loci with large effects detected in a human family were enriched with rare regulatory variants (44). A burden test for enrichment revealed a significant excess of rare regulatory variants at both extremes of gene expression, implicating their potential role in contributing to disease by driving high or low transcription (45). Our method can contribute to the identification and prioritization of rare variants.

DATA AVAILABILITY

The following repository contains source codes for feature set construction and CNN model training: https://github.com/kaistomics/cnnGWAS. Resources for model construction are also available at: https://omics.kaist.ac.kr/resources.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Brain Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT [2017M3C7A1048092]; Development of Health Prediction Technology Based on Big Data [KSN1713092] of the Korea Institute of Oriental Medicine. Funding for open access charge: Ministry of science and ICT [2017M3C7A1048092].
Conflict of interest statement. None declared.

REFERENCES

1. Edwards,S.L., Beesley,J., French,J.D. and Dunning,M. (2013) Beyond GWASs: Illuminating the dark road from association to function. *Am. J. Hum. Genet.*, **93**, 779–797.
2. Furey,T.S. and Sethupathy,P. (2013) Genetics driving epigenetics. *Science*, **342**, 705–706.
3. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
4. Farh,K.K., Marson,A., Zhu,J., Kleinewietfeld,M., Housley,W.J., Beik,S., Shoresh,N., Whitton,H., Ryan,R.J.H., Shishkin,A.A. *et al.* (2014) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.
5. Gjoneska,E., Pfenning,A.R., Mathys,H., Quon,G., Kundaje,A., Tsai,L.-H. and Kellis,M. (2015) Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature*, **518**, 365–369.
6. LeCun,Y., Bengio,Y. and Hinton,G. (2015) Deep learning. *Nature*, **521**, 436–444.
7. Xiong,H.Y., Alipanahi,B., Lee,L.J., Bretschneider,H., Merico,D., Yuen,R.K.C., Hua,Y., Gueroussov,S., Najafabadi,H.S., Hughes,T.R. *et al.* (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science*, **347**, 1254806.
8. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
9. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
10. Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, **381**, 1371–1379.
11. Lee,S.H., Ripke,S., Neale,B.M., Faraone,S. V, Purcell,S.M., Perlis,R.H., Mowry,B.J., Thapar,A., Goddard,M.E., Witte,J.S. *et al.* (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.*, **45**, 984–994.
12. Okada,Y., Wu,D., Trynka,G., Raj,T., Terao,C., Ikari,K., Kochi,Y., Ohmura,K., Suzuki,A., Yoshida,S. *et al.* (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.
13. Bentham,J., Morris,D.L., Graham,D.S.C., Pinder,C.L., Tombleson,P., Behrens,T.W., Martin,J., Fairfax,B.P., Knight,J.C., Chen,L. *et al.* (2015) Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.*, **47**, 1457–1464.
14. Liu,J.Z., van Sommeren,S., Huang,H., Ng,S.C., Alberts,R., Takahashi,A., Ripke,S., Lee,J.C., Jostins,L., Shah,T. *et al.* (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, **47**, 979.
15. The 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
16. Pasaniuc,B., Zaitlen,N., Shi,H., Bhatia,G., Gusev,A., Pickrell,J., Hirschhorn,J., Strachan,D.P., Patterson,N. and Price,A.L. (2014) Fast

and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**, 2906–2914.

17. The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
18. Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
19. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
20. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
21. Matys,V., Fricke,E., Geffers,R., Gössling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O. V *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
22. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
23. Zhang,C. and Ma,Y. (2012) *Ensemble Machine Learning*. Springer.
24. Liaw,A. and Wiener,M. (2002) Classification and regression by randomForest. *R News*, **2**, 18–22.
25. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
26. Fernández,J.M., de la Torre,V., Richardson,D., Royo,R., Puiggròs,M., Moncunill,V., Fragkogianni,S., Clarke,L., Flicek,P., Rico,D. *et al.* (2016) The BLUEPRINT data analysis portal. *Cell Syst.*, **3**, 491–495.
27. Sullivan,P.F., Daly,M.J. and O'Donovan,M. (2012) Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat. Rev. Genet.*, **13**, 537–551.
28. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
29. Sung,M.K., Jang,J., Lee,K.S., Ghim,C.-M. and Choi,J.K. (2016) Selected heterozygosity at cis-regulatory sequences increases the expression homogeneity of a cell population in humans. *Genome Biol.*, **17**, 164.
30. Rentzsch,P., Witten,D., Cooper,G.M., Shendure,J. and Kircher,M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
31. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W. and Richards,S. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
32. Huang,Y.-F., Gulko,B. and Siepel,A. (2017) Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.*, **49**, 618–624.
33. Yarwood,A., Huizinga,T.W.J. and Worthington,J. (2015) The genetics of rheumatoid arthritis: Risk and protection in different stages of the evolution of RA. *Rheumatol. (United Kingdom)*, **55**, 199–209.
34. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
35. Martin,P., McGovern,A., Orozco,G., Duffus,K., Yarwood,A., Schoenfelder,S., Cooper,N.J., Barton,A., Wallace,C., Fraser,P. *et al.* (2015) Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.*, **6**, 10069.
36. Kasela,S., Kisand,K., Tserel,L., Kaleviste,E., Remm,A., Fischer,K., Esko,T., Westra,H.J., Fairfax,B.P., Makino,S. *et al.* (2017) Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+versus CD8+T cells. *PLoS Genet.*, **13**, e1006643.
37. Plagnol,V., Smyth,D.J., Todd,J.A. and Clayton,D.G. (2009) Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics*, **10**, 327–334.
38. Yang,D., Jang,I., Choi,J., Kim,M.S., Lee,A.J., Kim,H., Eom,J., Kim,D., Jung,I. and Lee,B. (2018) 3DIV: a 3D-genome interaction viewer and database. *Nucleic Acids Res.*, **46**, D52–D57.
39. English,J.A., Fan,Y., Föcking,M., Lopez,L.M., Hryniewiecka,M., Wynne,K., Dicker,P., Matigian,N., Cagney,G., Mackay-Sim,A. *et al.* (2015) Reduced protein synthesis in schizophrenia patient-derived olfactory cells. *Transl. Psychiatry*, **5**, e663.
40. Wei,H., Yuan,Y., Liu,S., Wang,C., Yang,F., Lu,Z., Wang,C., Deng,H., Zhao,J., Shen,Y. *et al.* (2015) Detection of circulating miRNA levels in schizophrenia. *Am. J. Psychiatry*, **172**, 1141–1147.
41. Dickson,S.P., Wang,K., Krantz,I., Hakonarson,H. and Goldstein,D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.
42. Bodmer,W. and Bonilla,C. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
43. Wang,K., Dickson,S.P., Stolle,C.A., Krantz,I.D., Goldstein,D.B. and Hakonarson,H. (2010) Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 730–742.
44. Li,X., Battle,A., Karczewski,K.J., Zappala,Z., Knowles,D.A., Smith,K.S., Kukurba,K.R., Wu,E., Simon,N. and Montgomery,S.B. (2014) Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.*, **95**, 245–256.
45. Zhao,J., Akinsanmi,I., Arafat,D., Cradick,T.J., Lee,C.M., Banskota,S., Marigorta,U.M., Bao,G. and Gibson,G. (2016) A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.*, **98**, 299–309.