PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Annotating functional effects of non-coding variants in neuropsychiatric cell types by deep transfer learning

Boqiao Lai[1☯], **Sheng Qian**[2☯], **Hanwei Zhang**[3], **Siwei Zhang**[3], **Alena Kozlova**[3], **Jubao Duan**[3,4], **Jinbo Xu**[1*], **Xin He**[2*]

1 Toyota Technological Institute at Chicago, Chicago, Illinois, United States of America, **2** Department of Human Genetics, University of Chicago, Chicago, Illinois, United States of America, **3** Center for Psychiatric Genetics, NorthShore University HealthSystem, Evanston, Illinois, United States of America, **4** Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
* xinhe@uchicago.edu (XH); j3xu@ttic.edu (JX)

Citation: Lai B, Qian S, Zhang H, Zhang S, Kozlova A, Duan J, et al. (2022) Annotating functional effects of non-coding variants in neuropsychiatric cell types by deep transfer learning. PLoS Comput Biol 18(5): e1010011. https://doi.org/10.1371/journal.pcbi.1010011

Editor: Tony Capra, University of California San Francisco, UNITED STATES

Received: July 16, 2021

Accepted: March 11, 2022

Published: May 16, 2022

Data Availability Statement: The ATAC-seq data of the iPSC derived neurons are publicly available in the Gene Expression Omnibus(GSE129017). The reference epigenomic dataset is available at (http://deepsea.princeton.edu). Detailed data source for the neurodevelopmental model used in our experiment can be found in the supporting files (Table A in S1 Table).

Funding: This research was supported by the National Institutes of Health(https://www.nih.gov/) (R01MH116281, R01MH110531 to X.H. and

Abstract

Genomewide association studies (GWAS) have identified a large number of loci associated with neuropsychiatric traits, however, understanding the molecular mechanisms underlying these loci remains difficult. To help prioritize causal variants and interpret their functions, computational methods have been developed to predict regulatory effects of non-coding variants. An emerging approach to variant annotation is deep learning models that predict regulatory functions from DNA sequences alone. While such models have been trained on large publicly available dataset such as ENCODE, neuropsychiatric trait-related cell types are under-represented in these datasets, thus there is an urgent need of better tools and resources to annotate variant functions in such cellular contexts. To fill this gap, we collected a large collection of neurodevelopment-related cell/tissue types, and trained deep Convolutional Neural Networks (ResNet) using such data. Furthermore, our model, called MetaChrom, borrows information from public epigenomic consortium to improve the accuracy via transfer learning. We show that MetaChrom is substantially better in predicting experimentally determined chromatin accessibility variants than popular variant annotation tools such as CADD and delta-SVM. By combining GWAS data with MetaChrom predictions, we prioritized 31 SNPs for Schizophrenia, suggesting potential risk genes and the biological contexts where they act. In summary, MetaChrom provides functional annotations of any DNA variants in the neuro-development context and the general method of MetaChrom can also be extended to other disease-related cell or tissue types.

Author summary

A large number of genetic variants have been statistically associated with the risks of common diseases. However, whether such variants are actual risk variants and when and where they function are often unknown. To address this challenge, machine learning

R01GM089753 to J.X.), and the university of Chicago Biological Sciences Division(https://biologicalsciences.uchicago.edu/) (BSD 2021-22 to S.Q.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

methods have been developed to predict functional variants in specific cellular contexts. These methods correlate DNA sequences with their biological functions, e.g. enhancer activities, and can predict effects of single base mutations. Nevertheless, the training data used by existing methods often lack neurodevelopment-related cell types, thus annotating variant effects in neuropsychiatric genetics remains difficult. In this work, we fill this gap by collecting a large set of regulatory genomic datasets from fetal and adult brain, from iPSC-based cellular models and brain organoids. We trained deep learning models on this data, and further improved its performance by borrowing information from large external datasets, a strategy known as transfer learning. Our tool, MetaChrom, is substantially better at predicting experimentally determined regulatory variants than current methods, and helps us identify candidate risk variants of Schizophrenia. We believe MetaChrom provides a valuable tool for the neuropsychiatric genetic community, and the software can be of interest to researchers in other fields as well.

Introduction

GWAS of neuropsychiatric traits have identified hundreds of associated loci [1], however, translating these associations into detailed molecular mechanisms remain difficult. Most of variants in these loci are located in non-coding regions of the genome, with limited functional information. This makes it difficult to identify causal variants and target genes [2, 3]. In parallel to GWAS using common variants, sequencing studies also uncovered an important role of rare non-coding variants in regulatory elements in autism [4] and developmental disorders [5]. Given the low allele frequencies and reduced power of studying rare variants, having limited functional information of non-coding variations poses an even bigger problem. A key challenge in neuropsychiatric genetics is thus better annotation of functional effects of non-coding variants, ideally in a cell-type and allele-specific fashion [6].

To this end, experimental scientists have generated, in brain and neuronal cells, epigenomic maps, including chromatin accessibility and various histone marks. These data, however, usually annotate regulatory elements of hundreds of base pairs, and do not provide functional annotations at the base-level resolution. Various machine learning methods have been developed to fill in this gap. One class of methods, e.g. CADD [7], GWAVA [8], use conservation and epigenomic features, to predict likely functional variants, often based on a training set of known pathogenic variants. This approach, however, generally cannot predict allelic effects and lacks single-base resolution. More importantly, the training sets are often limited, as a result, these methods usually predict a general index of "pathogenicity" instead of context-specific effects. Another class of methods directly predict epigenomic profiles, such as protein binding sites, chromatin accessibility, histone marks and methylation, from DNA sequences [9–15]. Once trained, these models can predict the regulatory effect of a DNA variant by comparing predicted epigenomic properties of different alleles [16, 17]. These sequence-based methods can obtain single-nucleotide, and allele-specific prediction of variant effects on epigenomic features in specific cellular contexts. Because of these advantages, this approach has received great attention in the past few years [18]. In particular, deep learning based methods, such as Convolutional neural networks (CNNs), outperform traditional machine learning models in sequence-based prediction of protein-DNA/RNA interaction and chromatin profiles [16–22].

Despite these successes, it remains challenging to annotate variant effects, which often vary with cell types/tissues, in specific phenotypic contexts. Pre-trained models using public

epigenomic datasets such as ENCODE and Roadmap Epigenomics Consortium [2, 23] may not include the cell types of interest and thus not able to provide the correct variant annotations in those cell types [24]. This is particularly the case for neuropsychiatric traits. Because of the difficulty of collecting samples from developing human brain, publicly available epigenomic datasets contain only a limited set of postmortem brain samples (usually adult) [25–27]. As a result, cell and tissue types relevant to early neurodevelopment, which is important for genetics of neuropsychiatric traits [25, 28, 29], are under-represented in the training set of most current variant analysis tools, making it difficult to annotate functions of variants during neurodevelopment.

We proposed to address this challenge, by collecting a large set of neurodevelopment related epigenomic datasets, while taking advantage of additional datasets with a deep transfer learning framework. We collected 31 datasets from both fetal and postmortem brains, and from cellular models of early neurodevelopment, including brain organoid and induced Pluripotent Stem Cell (iPSC) derived neuronal cells [30]. Regulatory sequences in these cellular models, as our recent work demonstrated, differ substantially from those in adult brains, and are enriched with risk variants of neuropsychiatric traits [29, 31]. Using these datasets, we trained deep Convolutional Residual Networks (ResNet). ResNet is a technique that may train very deep CNNs to enhance the predictive power, and has been proven effective in computational biology problems such as RNA binding motif discovery [32] and protein folding [33]. To further improve the performance of Resnet, we use transfer learning, a general machine learning approach that leverages knowledge and models gained from one domain to a related domain [34, 35]. Specifically, we use a CNN-based meta-feature extractor to learn rich sequence features from the 919 external epigenomic profiles of diverse cell and tissue types [2, 23], and then combine them with ResNet to learn a sequence model for the neurodevelopmental epigenomic datasets. Our strategy thus has the advantage of rich representation of ResNet, while avoiding overfitting by using sequence features learned from external datasets.

Our approached, called MetaChrom, outperforms previous deep learning methods [16] and models without transfer learning, in predicting epigenomic profiles of our data. These higher predictive accuracy translates to a better prediction of functional single nucleotide variants, as measured by their effects on chromatin accessibility. We leverage this ability of MetaChrom to annotate likely effects of variants to study genetics of schizophrenia (SCZ), a complex mental disorder. The risk of SCZ has been associated with more than 100 genetic loci via GWAS, but in most loci, the causal variants remain unknown [1]. Combining neurodevelopment-specific predictions of variant effects by MetaChrom with GWAS results, we highlight 31 likely functional Single Nucleotide Polymorphism (SNPs) in 30 SCZ-associated loci. Studying these variants points to putative causal genes in these loci and the cell types and developmental stages in which these variants likely act.

Results

MetaChrom: Sequence-based prediction of epigenomic profiles and variant effects using transfer learning

We have built a general deep learning model to annotate regulatory variant effects, using only DNA sequences, with limited training data. Our training set consists of a set of DNA sequences, 1000 bps in length, and their functional labels, e.g. whether a sequence is in open chromatin region or not, in a given cell/tissue type. Additionally we have access to a large compendium of publicly available epigenomic profiles—the reference epigenomic data, which will be used to extract sequence features to improve model learning capability. The modular framework we have built, MetaChrom has two major components (Fig 1A): (1) a meta-feature

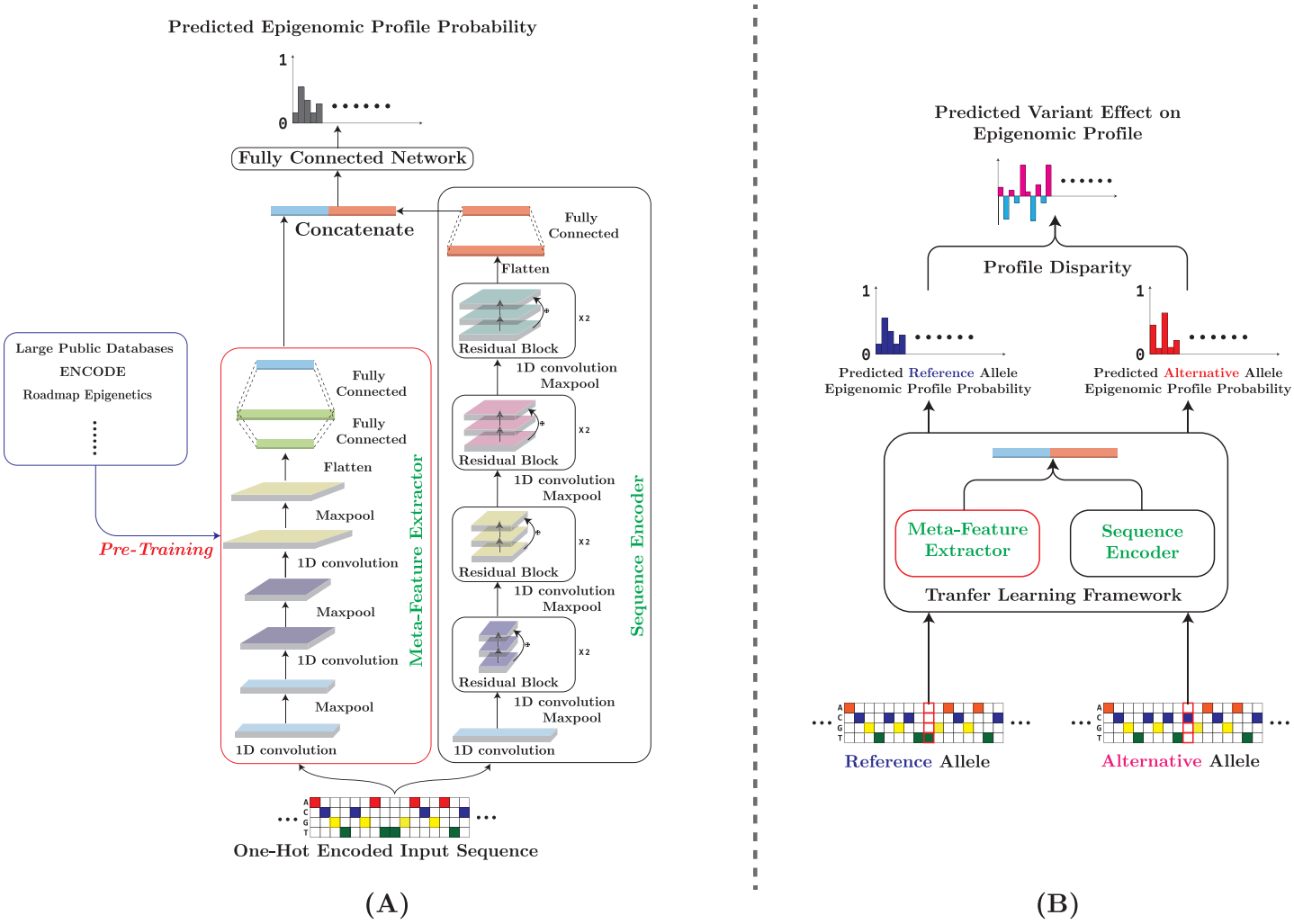

Fig 1. (A) Overall architecture of MetaChrom. The input sequence is fed into both MetaFeat and the ResNet sequence encoder. Their outputs are then concatenated for the prediction of epigenomic profiles.(B) Pipeline for predicting variant effect on sequence epigenomic profiles.

https://doi.org/10.1371/journal.pcbi.1010011.g001

extractor (MetaFeat) pre-trained on the reference epigenomic data; (2) a ResNet based sequence encoder. The meta-feature and the encoded sequence are then combined to predict the epigenomic profiles of the sequence of interest. The meta-feature, i.e. transfer learning, component learns important sequence features from the reference set. While precise interpretation of sequence features in deep neural networks is generally difficult, conceptually these sequence features can be viewed as certain "regulatory code", e.g. TF binding motifs or synergistic interaction between pairs of motifs. As shown later, learning such features would improve the model performance. Once a sequence-to-function model is trained, MetaChrom will be able to predict regulatory effects of genomic variants (Fig 1B), by comparing the predicted functional labels of two sequences differing in a single nucleotide.

MetaChrom accurately predicts epigenomic profiles across neurodevelopment-related cell types

We applied MetaChrom to predict 31 epigenomic features, including chromatin accessibility and histone marks of enhancers, derived from both fetal and adult brain tissues or neuronal

cells (See Method 4.2, Table A in S1 Table). The test sequences were obtained from chromosome 7 and 8, which were not used in the training process. We compared MetaChrom with other deep learning based methods in literature for predicting epigenomic profiles from DNA sequences. We note that the architecture of those models may depend on specific training data and it may not be easy to directly compare them with MetaChrom. For example, in the case of DeepSEA, it was designed for a large set of 919 epigenomic datasets. We thus implemented a baseline CNN model (BaseCNN) with 3 convolutional layers as representative of CNN based methods such as DeepSEA and Basset [16, 17]. We are also interested in the question of whether average epigenomic activities of a sequence across a large collection of cell types would be a good predictor of its activity in a new cell type. Such possibility has been raised in several recent papers [36, 37]. We thus obtained average epigenomic profiles, from DeepSEA, across a broad range of 919 cell and tissue types/conditions, denoted as DeepSEA-average.

We evaluated the performance of these methods and MetaChrom in predicting sequence labels in the testing data using Area under Precision-Recall curve (AUPRC) and Area under Receiver Operating Characteristic (AUROC) curve. Across the 31 cell-types of interest, MetaChrom achieved higher average AUROC (0.90), and AUPRC (0.53), than the BaseCNN model and DeepSEA-average (Fig 2). Our results show the advantage of MetaChrom comparing with standard CNN and off-the-shelf tools not trained for specific cell types of interest. To give a detailed picture of how the models perform, we showed the Response-Operating Curves and Precision-Recall curves of the three methods in Amygdala neurons in Fig A in S1 Text and the complete results in Figs B and C in S1 Text.

To further investigate the importance of transfer learning and the contribution of model architecture (ResNet vs. CNN) to the performance, we performed additional comparison of MetaChrom against two variants of MetaChrom: one without transfer learning and one where CNN instead of ResNet is used. When transferred knowledge is not used, our ResNet has average AUPRC = 0.28 and average AUROC = 0.80 across 31 cell types. ResNet with the meta-

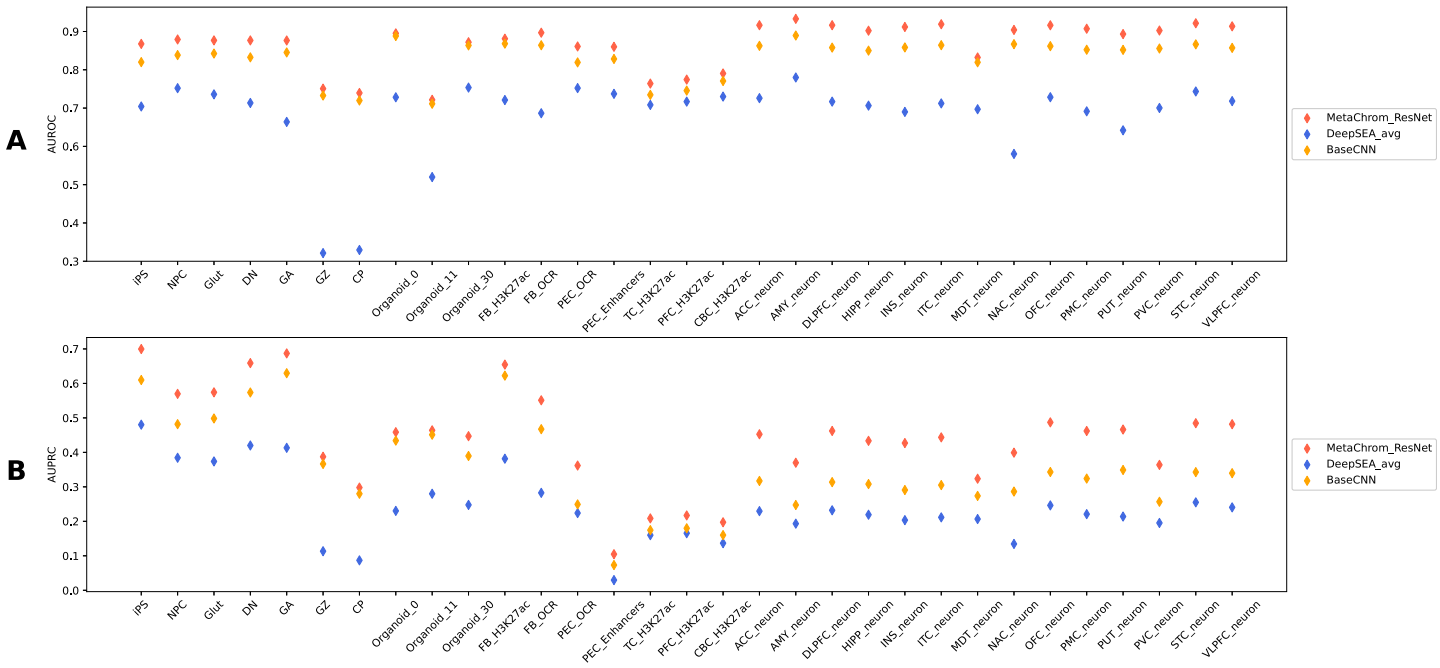

Fig 2. (A) AUROC and (B) AUPRC performance comparison of MetaChrom and DeepSEA method across 31 epigenomic features. NPC, Glut, DN, GA: iPSC-derived neurons. GZ, CP: germinal zone and cortical plate. OCR: open chromatin regions. See Table A in S1 Table for the list of cell/tissue types.

https://doi.org/10.1371/journal.pcbi.1010011.g002

feature extractor dramatically improves the performance: increasing its AUPRC from 0.28 to 0.50 and AUROC from 0.80 to 0.89, as shown in Figs D and E in S1 Text.

Our results thus highlight the advantage of the transfer learning approach of MetaChrom. We notice that our evaluation uses a large collection of 31 epigenomic features. We hypothesize that the advantage of MetaChrom over CNN would be even larger with smaller training set. This is likely a more common scenario in practice when a researcher trains a model for specific ell types of interest. To test this, we train MetaChrom and CNN on ATAC-seq data from iPSC and four types of iPSC-derived neurons. Across the five tested cell types, MetaChrom yielded AUROC of 0.87 and AUPRC of 0.82 while the average DeepSEA predictions yielded AUROC of 0.63 and AUPRC of 0.57 as shown in Fig A in S1 Text.

In summary, we demonstrate that MetaChrom is a powerful framework of predicting epigenomic profiles from DNA sequences, outperforming existing methods. Its power lies in both its ResNet architecture and its ability of transfer learning from external datasets.

MetaChrom predicted functional variants are supported by evolutionary constraint and allelic effects on chromatin accessibility

Evolutionary constraint is a commonly used metric of functional sequences [38]. We thus evaluated the accuracy of MetaChrom in predicting functional DNA variants by assessing evolutionary constraint on MetaChrom predicted variants. For all SNPs within peak regions of epigenomic data of each cell type, we computed MetaChrom scores, defined as the absolute value of the difference of MetaChrom predictions between reference and alternative alleles (Fig 1B). From these SNPs, we chose top 10,000 as predicted functional variants, and randomly sampled 100,000 variants from peak regions in the same cell type as control. We compared GERP scores, a commonly used measure of inter-species conservation [39], between predicted functional and control SNPs. In most cell types, MetaChrom top variants have significantly higher GERP scores than random ones (Fig 3A for a subset of cell types, the rest in Fig F in S1 Text), suggesting stronger evolutionary constraint. These results were confirmed with Human PhyloP scores from large 241-way mammalian alignment [40] (Fig G(A) in S1 Text). Given that functional sequences in brain may evolve relatively recently, we also assess the

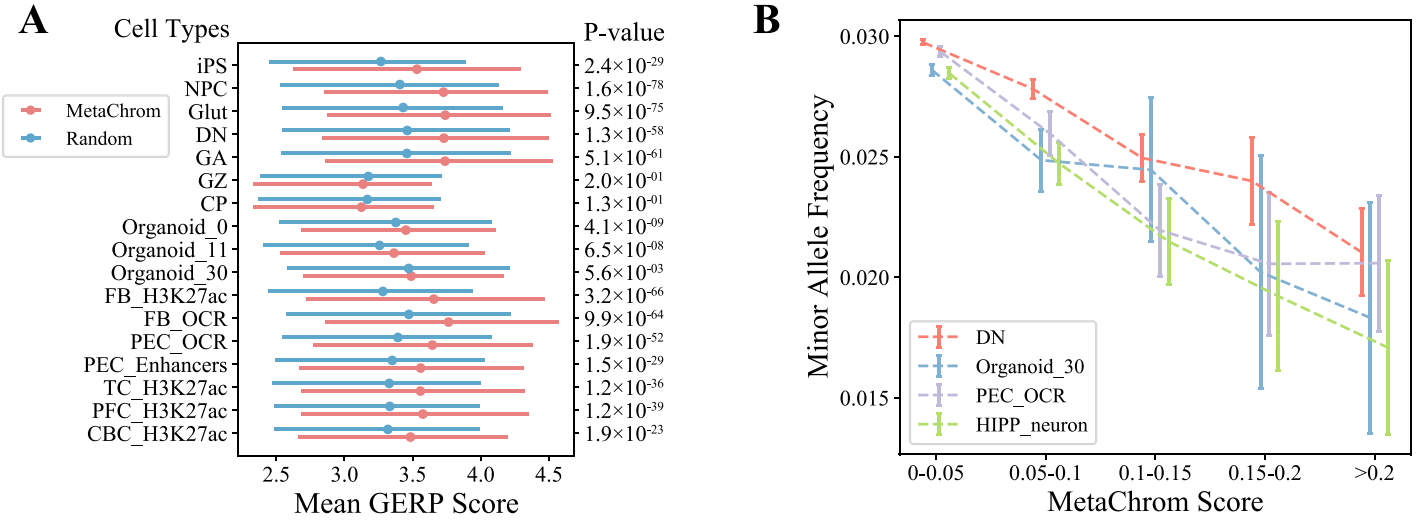

Fig 3. Validation of MetaChrom predicted functional variants with evolutionary constraint. (A) Distribution of GERP scores between MetaChrom predicted functional variants and random variants. (B) Minor allele frequencies of variants defined by MetaChrom scores in four selected cell types. Only variants inside peak regions of the epigenomic data were considered.

https://doi.org/10.1371/journal.pcbi.1010011.g003

evolutionary constraint using only primate genomes. This analysis shows similar results (Fig G in S1 Text). We next evaluated intra-species constraint of MetaChrom variants. Because of purifying selection, functionally deleterious variants often occur at low frequencies in the population [41, 42]. We obtained minor allele frequencies (MAFs) from the gnomAD database of all variants within peak regions of 31 epigenomic profiles. We observed a clear negative correlation between MAFs and MetaChrom scores, with high scoring variants present at lower MAFs (Fig 3B for two fetal and two adult cell types, the complete results in Figs H and I in S1 Text). This results thus support the deleterious effects of MetaChrom predicted functional variants.

To further validate MetaChrom, we compare its predictions with experimentally determined regulatory variants in iPSC-derived neurons, based on allele-specific chromatin accessibility (ASC) analysis [29]. ASC variants are defined by allelic imbalance in ATAC-seq experiments, reflecting allelic effects on chromatin accessibility and potentially gene expression. These ASC variants in iPSC-derived neurons were enriched with variants associated with gene expression, histone modification, DNA methylation, and neuropsychiatric traits [29]. We focused on ASC variants from neural progenitor cells (NPC) and glutamatergic (iN-Glut) neurons, two cell types with largest numbers of identified ASC variants. For all common single nucleotide variants (SNVs) in open chromatin regions of these two cell types, we computed their MetaChrom scores trained from the matched cell types. The top ranked 1,000 variants show about 6 fold enrichment of ASC variants, comparing with randomly sampled variants in open chromatin regions (Fig 4A). We also observed that MetaChrom scores from matched cell types generally show higher enrichment than scores from other cell types, confirming the cell type specificity of MetaChrom predictions (Fig J in S1 Text). For comparison, we also ranked variants within open chromatin regions by four other tools, including CADD, deltaSVM, FunSig, as well as a baseline CNN model trained on our collection of 31 epigenomic datasets [7, 10, 16]. CADD is widely used to predict deleteriousness of variants using a combination of evolutionary and epigenomic features. deltaSVM is a Support Vector Machine (SVM)-based supervised model for predicting variant effects, trained on the ATAC-seq data of the target cell types. Funsig is an aggregate measure of predicted regulatory effects, based on DeepSEA predictions from a large compendium of cell/tissue types (most are not from brain). Top variants by all these methods show varying levels of enrichment in ASC, but at levels lower than

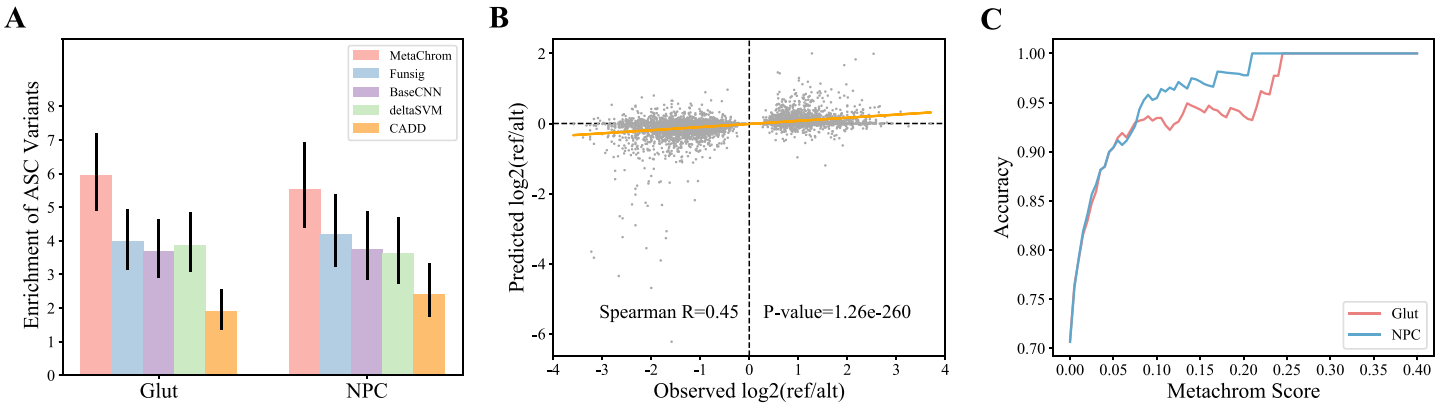

Fig 4. Validation of MetaChrom predicted functional variants with ASC variants. (A) Enrichment of ASC variants for predicted functional variants identified by MetaChrom, Funsig, CADD, deltaSVM, baseline CNN score in iN-Glut and NPC cells. (B) The observed allelic imbalance vs. MetaChrom predicted effects on chromatin accessibility of ASC variants in Glut neurons. (C) Accuracy of predicting directions of ASC variants in Glut and NPC cells.

https://doi.org/10.1371/journal.pcbi.1010011.g004

MetaChrom (Fig 4A). These results are robust to the number of top variants, and evaluation using Precision-Recall curve shows similar results (Fig K in S1 Text).

To test if MetaChrom can predict the effect sizes and directions of variants on chromatin accessibility, we compared the observed allelic imbalance of ASC variants in NPC and iN-Glut with the predicted differences between reference and alternative alleles. MetaChrom predictions track the observed allelic imbalance ratio with Spearman correlations of 0.45 and 0.40 in two cell types, respectively (Fig 4B and Fig L in S1 Text). Focusing on iN-Glut cells, we found 70% ASC variants show consistent signs in observed allelic imbalance and estimated effects (Fig 4B). ASC variants that are predicted to have large effects by MetaChrom show even higher agreement of predicted and observed directions of allelic imbalance. At MetaChrom score > 0.05, the agreement reaches nearly 90%, and goes even higher with higher MetaChrom score cutoff (Fig 4C). Together these results show that MetaChrom provides reasonable predictions of the regulatory effects of genetic variants.

MetaChrom assists interpretation of GWAS results

A single locus associated with a trait from GWAS could harbor hundreds of variants in linkage disequilibrium (LD), making it difficult to distinguish causal from non-causal variants. Recent work, including our own, have demonstrated that causal signals are enriched with variants disrupting chromatin states [29, 43, 44]. Motivated by this observation, we use MetaChrom to predict functional effects and identify putative causal variants of 145 SCZ-associated loci [45] (Table B in S1 Table). We score all the common SNPs by the absolute differences of MetaChrom predictions between two alleles in each of the 31 cell types we study. Additionally, we take into account the evidence of SNPs being causal variants from previous statistical fine-mapping analysis [45]. This analysis has identified candidate SNPs at each locus, known as credible set, and quantified the evidence of individual SNPs by Posterior Inclusion Probability (PIP), Bayesian posterior probability that a SNP is a causal variant given the GWAS data. We then combine the PIP values with MetaChrom scores to prioritize putative SCZ causal variants.

We identified 31 candidates in 30 SCZ-associated loci, based on several criteria: (i) plausibility of being SCZ risk variants (PIP > 0.1), (ii) MetaChrom scores ranked at top 1% across all common SNPs in at least one cell type, and (iii) MetaChrom scores are the very top among all SNPs in the credible set of a given SCZ risk locus, in at least one cell type (Fig 5). The list includes several high confidence SNPs with PIP > 0.5. Our results thus provide further support of the disease relevance of these SNPs, and additional information about how they may function, in terms of the relevant cell types and the epigenomic features they target. The majority of SNPs have moderate PIP values (0.1 to 0.5), and would not be considered causal SNPs by themselves. Using cell type specific MetaChrom scores, we can learn the biological context through which these variants work. Based on the cell types in which the MetaChrom score of a candidate SNP is highest among all SNPs in the credible set, we classify a SNP as acting likely in fetal stage (F) or in adult stage (A) or both (FA). Roughly equal number of SNPs in our candidates are classified as F or A, and six SNPs as FA. These findings are consistent with a recent report that expression associated variants of fetal and adult brain make comparable contributions to SCZ heritability [28]. Interestingly, once the stage (F or A) is given, the MetaChrom scores are often not very cell type specific. Most variants acting on adult stage have high scores across multiple types of adult neurons (Fig 5).

Even when causal variants are identified, their target genes may not be clear because of possible long-range regulation. To assign putative target genes, we leverage brain expression quantitative trait loci (eQTL) from post-morterm brain in GTEx [46] and CommonMinds

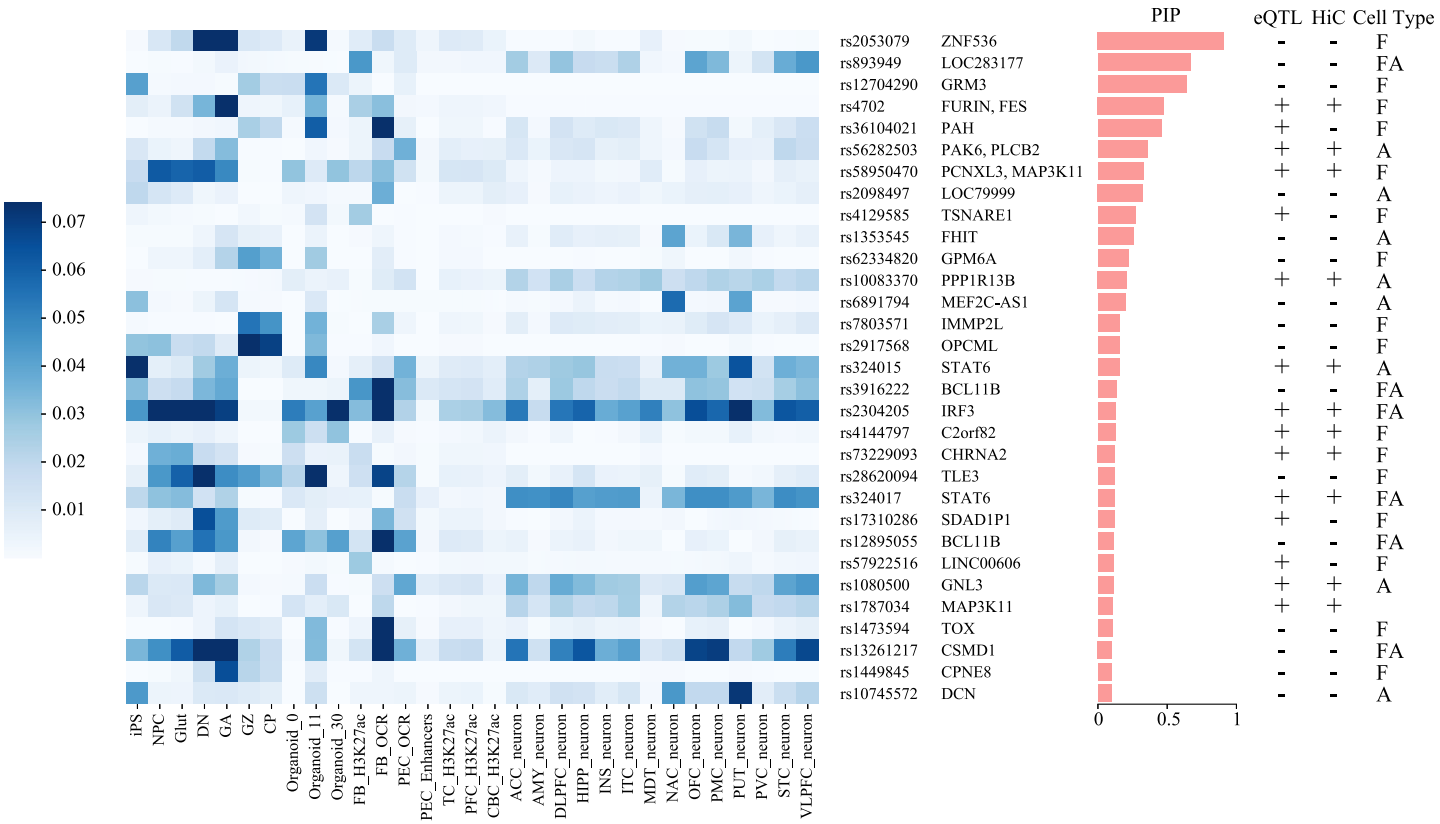

Fig 5. MetaChrom score of 31 candidate SNPs across 31 cell types. Candidate SNPs are ordered by their posterior inclusion probability (PIP) values shown in the middle. Three columns on the right indicate if a SNP is an eQTL (+/-), if a SNP has HiC targets (+/-) and if a SNP acts mostly in fetal stage (F) or in adult stage (A) or both (FA).

https://doi.org/10.1371/journal.pcbi.1010011.g005

Consortium (CMC) [47], and promoter-capture Hi-C data in iPSC-derived neurons [48]. A large fraction of our SNPs can be associated with one or more genes in eQTL, Hi-C or are located in the promoter and UTR regions. Combining these evidences and literature search, we assign the most likely target genes at each of the 31 SNPs (Fig 5 and Table C in S1 Table). Some these genes represent highly plausible risk genes of SCZ. For instance, FURIN and TSNARE1 were shown to regulate neuron growth and synaptic development by CRISPR editing in iPSC derived neurons [49]. GRM3 is a glutamate receptor and is being explored as a therapeutic target of SCZ [50]. ZNF356 is a transcription factor with an essential role in development of a subset of forebrain neurons implicated in stress and social behavior [51].

We discuss the SNP, rs2304205, in depth to show how MetaChrom may assist the study of genetics of complex traits (Fig 6). The region containing the SNP is strongly associated with SCZ, with multiple SNPs having *p*-values below genomewide threshold (Fig 6, top). Statistical fine-mapping is insufficient to resolve the causal variant in this locus. The credible set contains 12 SNPs, but the maximum PIP is below 0.2, suggesting the uncertainty of causal variants. MetaChrom analysis highlights rs2304205 as the most plausible causal variant. It has high scores across almost all cell types, in both fetal and adult stages (Fig 5). In 24/31 cell types we examined, rs2304205 has highest MetaChrom scores among the SNPs in the credible set (see four of these cell types, two each in fetal and adult stages, in Fig 6). The SNP is located in the UTR regions of both IRF3 and BCL2L12. Only IRF3 is found as the likely target gene of rs2304205 in brain eQTL data (Table C in S1 Table). IRF3 is a key regulator of the innate

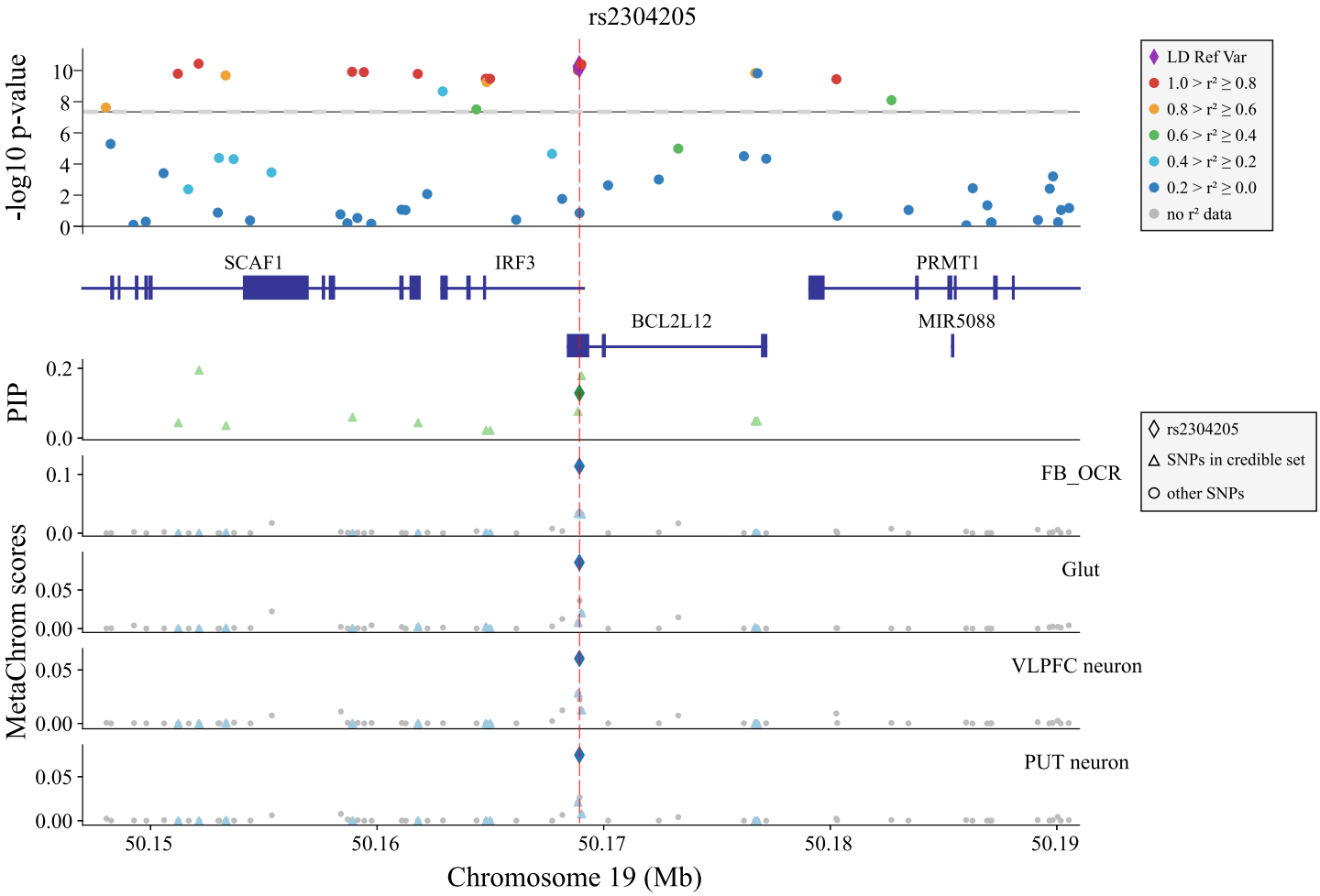

Fig 6. Likely causal variant rs2304205 and its MetaChrom functional annotations. The candidate SNP rs2304205 is chose as the reference variant for computing LD and it is highlighted by the red dash line in each panel. The upper panel shows significance of GWAS SNPs, LD between SNPs and genes in this region. The next panel shows credible set SNPs identified by fine-mapping (PIPs) in this region. The remaining panels show MetaChrom scores in four cell types, two in fetal stage (FB OCR and Glut) and two in adult stage (VLPEC neuron and PUT neuron).

https://doi.org/10.1371/journal.pcbi.1010011.g006

immune system [52]. Recent studies show that it may be important in regulating the development of neuronal progenitor cells [53], and physically interacts with other schizophrenia susceptibility genes, such as CREB1, AKT1 and ESR1 [52]. As another example, we performed additional study of a region containing rs1080500, another SNP highlighted by MetaChrom. The SNP effect is likely limited to adult neurons and is also an eQTL in adult brain (Fig M in S1 Text). The target gene based on eQTL, GNL3, is known to regulate neuron differentiation [54]. Taken together, these case studies highlight the potential of MetaChrom in prioritization of putative causal variants.

Identification of biologically relevant motifs from MetaChrom

To better understand what have been learned by our model, we extracted representative sequence patterns from our model using TF-MoDISco [55] and TOMTOM [56] (See Method 4.9). To focus on cell-type specific sequence features, we first sampled DNA sequence bins from our test set with mutually exclusive epigenomic feature labels. That is, each sampled

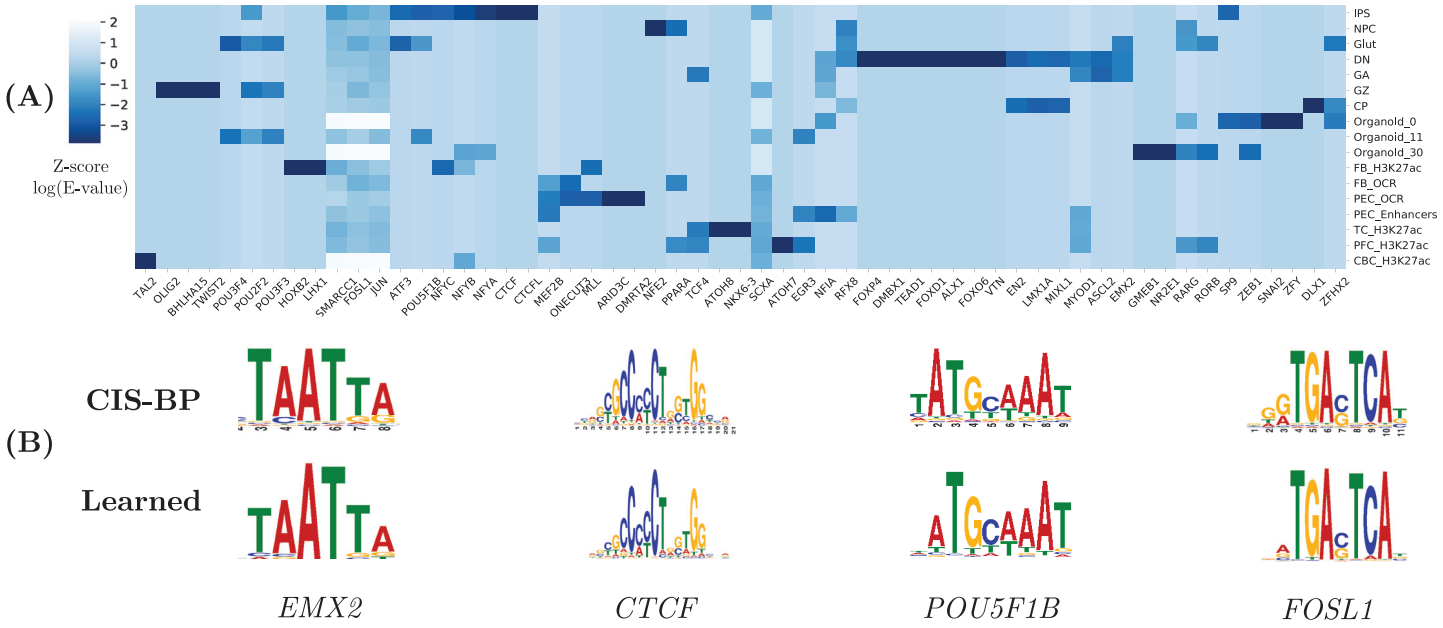

Fig 7. (A)Distribution of matched motifs in each epigenomic assay and (B)selected binding motifs identified by our method and their matches in the CIS-BP database.

https://doi.org/10.1371/journal.pcbi.1010011.g007

sequence fragment is marked positive in only one of the 31 epigenomic features. Then we computed the gradient with respect to the input sequence for cell-type-specific saliency signals and used it as the input to TF-MoDISco. The sequence patterns generated by TF-MoDISco are then matched to known TF (transcription factor) binding motifs at human CIS-BP [57] using TOMTOM [56].

As shown in Fig 7, our model detected known TF binding motifs specific to certain epigenomic features as well as motifs shared by multiple epigenomic features. For example, we detected binding motifs of the FOS and JUN families that form the activator protein 1 (AP-1) complex across various epigenomic features. These protein and protein complex are known to be associated with brain development [58, 59]. We detected many iPSC-specific motifs such as CTCF, an important regulator for chromatin structure [60], and POU5F1B(OCT4-PG1), a key player in the stem cell induction process [61]. In dopaminergic (DN) cell, we found motifs of proneural transcription factor FOXP4 that plays an important role in neural development [62] and FOXO6, a transcription factor closely related to cortical development [63, 64]. We discovered OLIG2, a transcription factor associated with cortical neurogenesis [65] and EN2 a transcription factor links to many stages of neural devlopment [66] in samples from the human neocortex(GZ, CP) [25].

Discussion

Computational prediction of functional non-coding variants can facilitate the discovery and interpretation of disease risk variants. In this paper we presented a deep transfer learning method, MetaChrom, combined with a large collection of epigenomic data from brain and neuronal cells, to predict regulatory effects of base-level DNA variants. By coupling ResNet [67] with transfer learning, we show that MetaChrom is accurate in predicting neurodevelopmental epigenomic profiles. Using a combination of evolutionary constraint and experimentally determined ASC variants, we validated the utility of MetaChrom in predicting functional

effects of variants. In particular, variants predicted by MetaChrom are substantially more enriched with ASC variants than the baseline CNN model and deltaSVM, two commonly used approach for annotating regulatory variants [43, 68]. We also illustrated how MetaChrom facilitates the prioritization of risk variants in GWAS loci associated with SCZ. The pre-trained models of MetaChrom are available online, and a user can query the likely functions of any variant(s) of interest using our server. We note that while the model was developed in the context of neurodevelopment, the approach and software is generic and can be applied to other user-provided epigenomic data in other biological contexts.

Many deep learning methods have been developed recently to study regulatory sequences and predict their effects. These models are generally based on CNN, e.g. DeepBind, DeepSEA, Basset, DeFine, and occasionally RNN (DanQ) [16, 17, 19, 22, 69]. It is not easy, however, to directly compare all these methods, as these models are optimized for specific training data. For instance, the popular DeepSEA method was trained on a larger set of 919 epogenomic features. For a fair comparison, we implemented an CNN model, which is currently the dominant architecture, and optimized the CNN model with our training data. As shown in our comparison, MetaChrom, by a combination of ResNet and transfer learning, outperforms the baseline CNN models in both epigenomic feature classification (Fig 2) and variant effect prediction (Fig 4).

One main application of deep learning based sequence models is the prediction of functional genetic variants. We demonstrated that MetaChrom predictions helped prioritize putative risk variants of SCZ and by combining with other datasets, revealed mechanistic insights of these variants. One limitation is that we do not yet have a single quantitative metric that combines statistical associations with deep learning based functional predictions. In our recent study [29], we show that it is possible to use ASC variants as functional information as prior in Bayesian fine-mapping. It would be interesting to extend such strategy to MetaChrom predictions. One possibility, for instance, is to combine MetaChrom predictions with experimental ATAC-seq data to have better power of detecting ASC variants. Such functionally informed genetic variant mapping has been used in eQTL studies and GWAS [70, 71]. The resulting set of deep learning-enhanced ASC set, when used as prior, may provide high resolution for fine-mapping GWAS loci.

In conclusion, we developed a deep learning based tool for predicting functional genetic variants. We demonstrated its accuracy using known regulatory variants in neuronal cells and its potential of revealing risk variants of GWAS of mental disorders. This tool is generally applicable and may enable researchers to better translate GWAS associations into mechanistic insights.

Materials and methods

Reference epigenomic profile data

We downloaded the epigenomic profile dataset from the DeepSEA website (http://deepsea.princeton.edu) [16]. This dataset consists of 919 chromatin features derived from the ENCODE and Roadmap Epigenomics [2, 72]. The epigenomic features are computed by first binning the reference genome (GRCh38/hg38) into 200-bp sequence fragments. The fragments were then intersected with the downloaded peaks from the public databases. Each bin was assigned a binary vector $l \in R^d (d = 919)$ as its label, each dimension representing an epigenomic feature i from a specific cell type with corresponding sequencing assay. If a fragment is at least 50% overlapped with the peak present in the sequencing assay i, the corresponding dimension in l is assigned 1 (i.e., $l_i = 1$), otherwise 0 (i.e., $l_i = 0$). After computing the features, the fragments were extended to 1kb to include surrounding sequences [16, 73]. All the

fragments were then split into training, validation, and test sets such that all fragments from chromosomes 7 and 8 are held out for test, and the rest are randomly split into training and validation sets. The training set we used contains 4,400,000 sequence fragments with at least one positive chromatin feature. We have trained our meta-feature extractor (MetaFeat) on this chromosomal-based split so we can fine-tune the feature extractor using the test sequence. We also employed the same chromosomal-based split in our case study for neurodevelopment related tissues data(Table A in S1 Table) which ensures the test sequences are never seen by our model before test time to avoid potential training bias.

Epigenomic profiles from neurodevelopment related tissues and cell types

To comprehensively capture the epigenomics landscape of the human neurons and brain, we collected data from 31 different epigenomic assays, from the early developmental stages to fully developed adult brain tissues. For the early developmental stages, we obtained a set of ATAC-seq peaks from iPSC derived neuronal cells described in [29], with a total of five different cell types. These cell types are good models of neurodevelopment. We also obtained one fetal brain DNase-seq sample from the Roadmap Epigenomics Project [23]; chromatin accessibility data from brain organoid samples at three different time points [74]; ATAC-seq profiles from two early human neocortex samples in germinal zone (GZ) and cortical plate (CP) [25]; and one fetal brain H3K27ac profile from [75]. For the adult brain, we collected fourteen neuronal ATAC-seq profiles from the BOCA project [76]; and five chromatin and histone features from the PsychENCODE project [77].

We processed the peaks from each epigenomic profile in a similar way as the aforementioned reference epigenomic profile data. In our dataset, each dimension in the label $l \in R^{31}$ represents if a segment is active or not in a given epigenomic profile, i.e., if it is at least 50% overlapped with the peaks in each epigenomic assay. We choose our test set such that all fragments from chromosomes 7 and 8 are held-out when the rest of the genome are randomly split for training and validation. After processing, we obtained 3,165,290 sequence fragments that's active in at least one epigenomic assay for training and validation; our test set contains 390,380 sequence fragments for model evaluation.

MetaChrom model architecture

MetaChrom as shown in Fig 1A has two major modules: 1) a CNN-based meta-feature extractor pre-trained on a large public dataset(MetaFeat). 2) a ResNet-based sequence model that learns cell-type-specific features directly from the input sequence. To predict the regulatory profile of a DNA sequence fragment i of 1kbp in length, we first encode the sequence fragment as a one-hot matrix and fed it into the meta-feature extractor, which will output a vector representation $\mathcal{F}_{meta}$ of the input sequence while the sequence is also fed to the ResNet-based sequence encoder simultaneously to obtain a cell-type-specific feature representation $\mathcal{F}_{seq}$. The two feature representations are concatenated to form a new joint vector representation $\mathcal{F}_{joint}$, which is fed into a fully connected dense network to predict epigenomic features $A \in R^{n_{out}}$, Where n_{out} is the number of epigenomic features of interest.

MetaFeat is a CNN model consisting of three 1-D convolutional layers with kernel length 7 and channel sizes 320, 480 and 960, respectively. This feature extractor takes a one-hot encoded sequence (of dimension 1000×4) as input and outputs a vector representation(R^{919}) of the input sequence fragment. Each convolutional layer in this feature extractor is followed by a ReLU [78] layer for activation and a max-pooling layer with a kernel size of 4 for down-sampling. We used a smaller kernel size (7), while previous methods use a large kernel size (19) for model interpretability [17], but it has been shown that with appropriate interpretation

tools [37, 55] meaningful binding motifs may be detected with a smaller kernel. The final convolutional layer connects to two fully connected layers, which in turn generate a vector of 919 chromatin features to represent the input.

Our ResNet-based sequence encoder maps the input sequence (encoded as a one-hot matrix of dimension 1000×4) to a cell-type-specific representation, which is a vector of 31 elements. The ResNet model consists of one 1D convolutional layer and eight residual blocks. The 1D convolutional layer has a kernel length 4 and channel size 48. Each residual block contains two 1D convolutional layers, each followed by a ReLU activation layer. The eight residual blocks have channel sizes of 96, 96 128, 128 256, 256, 512 and 512, respectively and kernel length 7. Finally, the last residual block connects to two fully connected layers, which generate the cell-type-specific representation for the input sequence segment.

The outputs from MetaFeat and the sequence encoder are concatenated to form a integrative vector representation of the input sequence, which is then fed into three fully connected layers to predict the probabilities of epigenomic profiles.

Other methods for comparison

To compare our proposed method with other alternatives in the community, we implemented a baseline CNN model(BaseCNN) with 3 convolutional layers [16] and trained the models on our curated dataset for comparison. We also compared our predictions with the mean epogenomic profile predicted by DeepSEA(DeepSEA_avg) for each test sequence which was directly obtained from their server [16]. CADD scores and funsig scores are downloaded from the respective server [7, 16]. See Fig L in S1 Text for the details.

Predicting variant effects on regulatory profiles by MetaChrom

To predict the variant effect on a given sequence's epigenomic profile, two sequences of length 1kb differing only at and centered at the variant position are used. One of them corresponds to the reference allele and the other corresponds to the alternative allele. As shown in Fig 1B, we pass those two sequences into MetaChrom separately to predict their epigenomic profiles as A_{ref} and A_{alt}. Then we compare the predicted regulator profile and compute the disparity as absolute difference $|A_{ref} - A_{alt}|$ or log odds ratio $\left|\log\left(\frac{A_{ref}}{1-A_{ref}}\right) - \log\left(\frac{A_{alt}}{1-A_{alt}}\right)\right|$ for measuring variant effects.

Model training and testing

We trained our transfer learning framework in two phases. In phase one we train the meta-feature extractor on the public epigenomic profile data (see Data section) using binary cross-entropy as the loss function and the Adam optimizer [79]. In phase two we jointly train the ResNet model and the meta-feature extractor [35] with stochastic gradient descent(SGD) accelerated by Nesterov momentum [80], starting from the trained feature extractor in the first phase with binary cross-entropy loss function on the targeted cellular context data. In both phases, sequence bins from chromosome 7 and 8 are held-out for test and the training/validation sets are randomly split from the remaining bins. For model selection, we searched batch sizes of (16, 32, 64, 128, 256), ten learning rates in (1e-5, 1e-3) equally divided in log scaled distance, as well as kernel sizes of (4,6,8) for the pre-training phase. For the joint training phrase, we searched on different batch sizes of (16, 32, 64, 128, 256), learning rates in (1e-5, 1e-3) equally divided with log scaled distance. We also tested different learning decay rates and Nesterov momentum. All models are trained on an NVIDIA 2080Ti GPU in 5 hours.

Assessing evolutionary constraint on MetaChrom predicted functional variants

A variant is scored by MetaChrom using the absolute value of the difference in MetaChrom output between reference and alternative alleles. The score ranges from 0 to 1. For top MetaChrom predicted variants in 31 cell types, we calculated and compared their GERP scores and PhyloP scores [39] with control variants, chosen randomly in peak regions of the same cell types. GERP scores were obtained from ANNOVAR [81], PhyloP scores were obtained from https://cglgenomics.ucsc.edu/data/cactus/ and http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phyloP17way/, and were compared between MetaChrom predicted and control variants using the Wilcoxon Rank-Sum Test.

Minor allele frequencies (MAFs) of all variants were obtained from the Genome Aggregation Database (gnomAD) [82] using ANNOVAR. We split variants into two sets of 5 bins based on the MetaChrom score: 0–0.05, 0.05–0.1, 0.1–0.15, 0.15–0.2, 0.2–1.0 and 0–0.05, 0.01–0.02, 0.02–0.03, 0.03–0.04, 0.04–1.0. In each bin, mean MAFs were calculated to investigate the correlation between MAFs and MetaChrom scores.

Evaluation of MetaChrom prediction using ASC variants

We used a recently published dataset of allele-specific chromatin accessibility (ASC) variants to compare several methods for predicting functional variants [29]. ASC variants are defined by allele imbalance in read counts from ATAC-seq experiments. A total of 5,611 and 3,547 ASC SNPs, at FDR < 0.05, were identified in neural progenitor cells (NPC) and gutamatergic neurons (iN-Glut), respectively.

To identify putative functional variants, we limit to single nucleotide variants (SNVs) in open chromatin regions in 2 cell types. The SNVs are retrieved from the 1000 Genomes Project with MAF $> 5\%$ [83].

We calculated scores of all these SNVs from several tools. Funsig scores were obtained from the DeepSEA Server [16] and CADD scores [7] were obtained from ANNOVAR. For deltaSVM [10, 84–86], we followed the training strategy from Shigaki et al [87]. Specifically, we trained gkm-SVM models on 300 bp sequences centered on ATAC-seq data of iN-Glut and NPC cells with LS-GKM. We set parameters -l(word length) to 11, -k(number of informative column) to 7, -d(maximum number of mismatches to consider) to 3, -t(kernal function) to 2, and other parameters follow default values. deltaSVM scores were then obtained with the trained gkm-SVM model and script deltasvm.pl. Software and scripts for deltaSVM can be found in http://www.beerlab.org/deltasvm/. Baseline CNN scores were obtained from the baseline CNN model that were trained on 31 epigenomic profiles. We chose the top 1,000 variants ranked by MetaChrom, Funsig, deltaSVM, CADD, baseline CNN score in descending order as predicted functional variants for each method. We then counted the number of ASC variants in predicted functional variants vs. control variants. The enrichment of ASC variants is then calculated by Fisher Exact Test.

In the analysis involving effect size and direction of SNPs, we define the observed allelic imbalance as $\log(R_{ref}/R_{alt})$, where R_{ref} and R_{alt} denote the number of reads mapped to the two alleles, and the MetaChrom predicted effects on chromatin accessibility from our model as $\log(A_{ref}/A_{alt})$. Correlation between observed allelic imbalance and MetaChrom predicted effects on chromatin accessibility is calculated by Spearman's rank correlation coefficient.

We also estimated the percent of MetaChrom predicted variants are actually experimentally determined ASC variants. For iN-Glut cells, among top 1000 MetaChrom variants, 462 SNPs are evaluated for allele imbalance test (not all SNPs are heterozygous in the study), and 123

(27%) are reported as ASC variants. For NPC, 445 SNPs, among top 1000, are evaluated for allele imbalance test and 87 (20%) are ASC variants.

Motif identification and visualization

Our model learns context-specific sequence patterns to predict epigenomic profiles. Many of these patterns may correspond to cell-type-specific transcription factor (TF) binding motifs. One way to interpret the models is to extract sequence patterns (motifs) from the filters of the first convolutional layer [17, 19, 69], but this strategy often yields patterns that are hard to interpret. This is because CNNs learn a distributed representation of sequence motifs and thus, an individual filter may correspond to only a partial motif that cannot be easily identified [55, 88]. Some methods assess the importance of each position in a given sequence to measure individual nucleotide contribution [89], but they do not yield interpretable motifs directly.

To extract meaningful sequence motifs from our deep model, here we used the recently-developed tool TF-MoDISco [55] that combines position-wise single-nucleotide contribution scores to generate cell-type-specific sequence patterns. We then use TOMTOM [56] to match the identified sequence patterns to known TF binding motifs in the CIS-BP database for further analysis [57]. To apply TF-MoDISco, we first randomly select 2,000 DNA bins from our testing set with mutually exclusive epigenomic features, in which each sequence bin is only marked active in one epigenomic feature but not others. Then we generate position-wise contribution score with saliency map [90], which is the gradient of the output with respect to the one-hot encoded input. The gradient is then gated by the observed nucleotide to generate importance score, i.e., only the gradient of the observed nucleotide is kept and the gradient of unobserved nucleotide is set to 0. The importance score is fed into TF-MoDISco to generate predictive sequence patterns, which were searched against the human CIS-BP [57] database for TF binding motifs using TOMTOM [56] with E-value = 1e-4.

Supporting information

S1 Text. Fig A. ROC and PRC plot for Amygdala neurons across different tested models. **Fig B**. (i) Average ROC and PRC plot for 31 epigenomic features across different tested models. (ii) Average ROC and PRC plot for five iPSC-derived neuronal cell types. The Average PRC performance appears better for the five cell-types model than the 31-cell-types model because some of the other cell types in the 31-cell-types model have low AUPRC values which resulted in lower average AUPRC. **Fig C**. ROC and PRC plot for (A) CNN, (B) ResNet, (C) MetaFeat-CNN, (D) MetaFeat-ResNet models on 31 epigenomic features. **Fig D**. Average ROC and PRC plot for 31 epigenomic features across different tested models. CNNBase and ResNet are baseline CNN and ResNet models without transfer learning. MetaFeat-CNN: CNN model with transfer learning. **Fig E**. (A) AUROC and (B) AUPRC performance comparison of MetaChrom and other methods across 31 epigenomic features. See Table A in S1 Table for the list of cell/tissue types. **Fig F**. Distribution of GERP scores between MetaChrom predicted functional variants and random variants sampled from the peak regions in each cell type. P-values testing the difference were computed from Wilcoxon Rank-Sum Test. **Fig G**. Evolutionary constraint evaluated by Human PhyloP scores (A) 241-way mammalian alignment from the Zoonomia Project (B) 17-way primate specific alignment. **Fig H**. Minor allele frequencies of variants defined by MetaChrom scores in 10 epigenomic profiles in fetal brain cell types. Only variants within peak regions of the data were considered. **Fig I**. Minor allele frequencies of variants defined by MetaChrom scores in 17 epigenomic profiles in adult brain cell types. Only variants within peak regions of the data were considered. **Fig J**. Number of experimentally determined ASC variants (two cell types, Glut—left and NPC—right) in top 10,000

MetaChrom predicted functional variants across 31 cell types. **Fig K**. Comparison of methods in predicting ASC variants (A) Zoomed-in Precision-Recall (PR) curve with Recall in [0,0.2] and Precision in [0,0.25]. (B) Number of ASC variants in top K prioritized variants. **Fig L**. The observed allelic imbalance vs. MetaChrom predicted effects on chromatin accessibility of ASC variants in NPC neurons. **Fig M**. Likely causal variant rs1080500 and its MetaChrom functional annotations. The candidate SNP rs1080500 is chosen as the reference variant for computing LD and it is highlighted by the red dash line in each panel. The upper panel shows the significance of GWAS SNPs, LD between SNPs and genes in this region. The next panel shows credible set SNPs identified by fine-mapping (PIPs) in this region. The remaining panels show MetaChrom scores in four cell types, two in the fetal stage (FB_OCR and Glut) and two in the adult stage (VLPFC neuron and OFC neuron). **Fig N**. Baseline CNN model Architecture.
(DOCX)

S1 Table. Table A. Cell type information. **Table B**. Creditable SNPs set. **Table C**. GWAS Candidate SNPs.
(XLSX)

Acknowledgments

We thank members of He and Xu labs for helpful discussions.

Author Contributions

Conceptualization: Boqiao Lai, Jinbo Xu, Xin He.

Data curation: Boqiao Lai, Sheng Qian, Hanwei Zhang, Siwei Zhang, Alena Kozlova, Jubao Duan.

Formal analysis: Boqiao Lai, Sheng Qian, Xin He.

Funding acquisition: Jinbo Xu, Xin He.

Investigation: Boqiao Lai, Sheng Qian, Xin He.

Methodology: Boqiao Lai, Sheng Qian, Jinbo Xu, Xin He.

Project administration: Jinbo Xu, Xin He.

Resources: Jinbo Xu, Xin He.

Software: Boqiao Lai.

Supervision: Jinbo Xu, Xin He.

Validation: Boqiao Lai, Sheng Qian, Xin He.

Visualization: Boqiao Lai, Sheng Qian.

Writing – original draft: Boqiao Lai, Sheng Qian, Jinbo Xu, Xin He.

Writing – review & editing: Boqiao Lai, Sheng Qian, Xin He.

References

1. Ripke S, Neale BM, Corvin A, Walters JT, Farh KH, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421–427. https://doi.org/10.1038/nature13595
2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57. https://doi.org/10.1038/nature11247

3. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337(6099):1190–1195. https://doi.org/10.1126/science.1222794 PMID: 22955828
4. An JY, Lin K, Zhu L, Werling DM, Dong S, Brand H, et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science. 2018; 362 (6420). https://doi.org/10.1126/science.aat6576 PMID: 30545852
5. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. De novo mutations in regulatory elements in neurodevelopmental disorders. Nature. 2018; 555(7698):611–616. https://doi.org/10.1038/nature25983 PMID: 29562236
6. Powell SK, O'Shea C, Brennand KJ, Akbarian S. Parsing the functional impact of noncoding genetic variants in the brain epigenome. Biological Psychiatry. 2021; 89(1):65–75. https://doi.org/10.1016/j.biopsych.2020.06.033 PMID: 33131715
7. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2018; 47(D1):D886–D894. https://doi.org/10.1093/nar/gky1016
8. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nature methods. 2014; 11(3):294–296. https://doi.org/10.1038/nmeth.2832 PMID: 24487584
9. Arnold P, Schöler A, Pachkov M, Balwierz PJ, Jørgensen H, Stadler MB, et al. Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. Genome research. 2013; 23 (1):60–73. https://doi.org/10.1101/gr.142661.112 PMID: 22964890
10. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nature genetics. 2015; 47(8):955. https://doi.org/10.1038/ng.3331 PMID: 26075791
11. Benveniste D, Sonntag HJ, Sanguinetti G, Sproul D. Transcription factor binding predicts histone modifications in human cell lines. Proceedings of the National Academy of Sciences. 2014; 111(37):13367–13372. https://doi.org/10.1073/pnas.1412081111 PMID: 25187560
12. Pinello L, Xu J, Orkin SH, Yuan GC. Analysis of chromatin-state plasticity identifies cell-type–specific regulators of H3K27me3 patterns. Proceedings of the National Academy of Sciences. 2014; 111(3): E344–E353. https://doi.org/10.1073/pnas.1322570111 PMID: 24395799
13. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, et al. Computational prediction of methylation status in human genomic sequences. Proceedings of the National Academy of Sciences. 2006; 103(28):10713–10716. https://doi.org/10.1073/pnas.0602949103 PMID: 16818882
14. Setty M, Leslie CS. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. PLoS computational biology. 2015; 11(5):e1004271. https://doi.org/10.1371/journal.pcbi.1004271 PMID: 26016777
15. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nature methods. 2014; 12(3):265. https://doi.org/10.1038/nmeth.3065 PMID: 25240437
16. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods. 2015; 12(10):931. https://doi.org/10.1038/nmeth.3547 PMID: 26301843
17. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research. 2016; 26(7):990–999. https://doi.org/10.1101/gr.200535.115 PMID: 27197224
18. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics. 2019; 20(7):389–403. https://doi.org/10.1038/s41576-019-0122-6 PMID: 30971806
19. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology. 2015; 33(8):831. https://doi.org/10.1038/nbt.3300 PMID: 26213851
20. Lam JH, Li Y, Zhu L, Umarov R, Jiang H, Héliou A, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nature communications. 2019; 10(1):1–13. https://doi.org/10.1038/s41467-019-12920-0 PMID: 31666519
21. Grønning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, et al. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. Nucleic acids research. 2020; 48 (13):7099–7118. https://doi.org/10.1093/nar/gkaa530 PMID: 32558887
22. Wang M, Tai C, E W, Wei L. DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic acids research. 2018; 46(11):e69–e69. https://doi.org/10.1093/nar/gky215 PMID: 29617928

23. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518(7539):317. https://doi.org/10.1038/nature14248 PMID: 25693563

24. Wesolowska-Andersen A, Yu GZ, Nylander V, Abaitua F, Thurner M, Torres JM, et al. Deep learning models predict regulatory variants in pancreatic islets and refine type 2 diabetes association signals. Elife. 2020; 9:e51503. https://doi.org/10.7554/eLife.51503 PMID: 31985400

25. de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, et al. The dynamic landscape of open chromatin during human cortical neurogenesis. Cell. 2018; 172(1-2):289–304. https://doi.org/10.1016/j.cell.2017.12.014 PMID: 29307494

26. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, et al. An atlas of chromatin accessibility in the adult human brain. Genome research. 2018; 28(8):1243–1252. https://doi.org/10.1101/gr.232488.117 PMID: 29945882

27. Bryois J, Garrett ME, Song L, Safi A, Giusti-Rodriguez P, Johnson GD, et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nature communications. 2018; 9(1):3121. https://doi.org/10.1038/s41467-018-05379-y PMID: 30087329

28. Walker R, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, Torre-Ubieta Ldl, et al. Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. Cell. 2019; 179(3):750–771. https://doi.org/10.1016/j.cell.2019.09.021 PMID: 31626773

29. Zhang S, Zhang H, Zhou Y, Qiao M, Zhao S, Kozlova A, et al. Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. Science. 2020; 369(6503):561–565. https://doi.org/10.1126/science.aay3983 PMID: 32732423

30. Vadodaria KC, Amatya DN, Marchetto MC, Gage FH. Modeling psychiatric disorders using patient stem cell-derived neurons: a way forward. Genome medicine. 2018; 10(1):1. https://doi.org/10.1186/s13073-017-0512-3 PMID: 29301565

31. Forrest MP, Zhang H, Moy W, McGowan H, Leites C, Dionisio LE, et al. Open chromatin profiling in hiPSC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. Cell stem cell. 2017; 21(3):305–318. https://doi.org/10.1016/j.stem.2017.07.008 PMID: 28803920

32. Koo PK, Anand P, Paul SB, Eddy SR. Inferring Sequence-Structure Preferences of RNA-Binding Proteins with Convolutional Residual Networks. bioRxiv. 2018; p. 418459.

33. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology. 2017; 13(1):e1005324. https://doi.org/10.1371/journal.pcbi.1005324 PMID: 28056090

34. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer; 2018. p. 270–279.

35. Li Z, Hoiem D. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence. 2017; 40(12):2935–2947. https://doi.org/10.1109/TPAMI.2017.2773081 PMID: 29990101

36. Schreiber J, Singh R, Bilmes J, Noble WS. A pitfall for machine learning methods aiming to predict across cell types. Genome biology. 2020; 21(1):1–6. https://doi.org/10.1186/s13059-020-02177-y PMID: 33213499

37. Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. bioRxiv. 2019; p. 605717. https://doi.org/10.1093/bioinformatics/btz352 PMID: 31510655

38. Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, et al. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Research. 2010; 20(3):301. https://doi.org/10.1101/gr.102210.109 PMID: 20067941

39. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Computational Biology. 2010; 6(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025 PMID: 21152010

40. Consortium Z. A comparative genomics multitool for scientific discovery and conservation. Nature. 2020; 587(7833):240–245. https://doi.org/10.1038/s41586-020-2876-6

41. Loewe L. Negative Selection. Nature Education. 2008; 1(1):59.

42. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):258–291. https://doi.org/10.1038/nature19057 PMID: 27535533

43. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, Lituiev D, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. Nature Genetics. 2018; 50(8):1140–1150. https://doi.org/10.1038/s41588-018-0156-2 PMID: 29988122

44. Calderon D, Nguyen MLT, Mezger A, Kathiria A, Müller F, Nguyen V, et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nature Genetics. 2019; 51(10):1494–1505. https://doi.org/10.1038/s41588-019-0505-9 PMID: 31570894

45. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature genetics. 2018; 50(3):381–389. https://doi.org/10.1038/s41588-018-0059-2 PMID: 29483656

46. Consortium G. Genetic effects on gene expression across human tissues. Nature. 2017; 550 (7675):204–213. https://doi.org/10.1038/nature24277

47. Fromer M, Roussos P, Sieberts SK, Johnson J, Kavanagh DH, Perumal TM, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience. 2016; 19(11):1442–1453. https://doi.org/10.1038/nn.4399 PMID: 27668389

48. Song M, Yang X, Ren X, Maliskova L, Li B, Jones IR, et al. Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. Nature Genetics. 2019; 51 (8):1252–1262. https://doi.org/10.1038/s41588-019-0472-1 PMID: 31367015

49. Schrode N, Ho SM, Yamamuro K, Dobbyn A, Huckins L, Matos MR, et al. Synergistic effects of common schizophrenia risk variants. Nature genetics. 2019; 51(10):1475–1485. https://doi.org/10.1038/s41588-019-0497-5 PMID: 31548722

50. García-Bea A, Walker MA, Hyde TM, Kleinman JE, Harrison PJ, Lane TA. Metabotropic glutamate receptor 3 (mGlu3; mGluR3; GRM3) in schizophrenia: Antibody characterisation and a semi-quantitative western blot study. Schizophrenia research. 2016; 177(1-3):18–27. https://doi.org/10.1016/j.schres.2016.04.015 PMID: 27130562

51. Thyme SB, Pieper LM, Li EH, Pandey S, Wang Y, Morris NS, et al. Phenotypic Landscape of Schizophrenia-Associated Genes Defines Candidates and Their Shared Functions. Cell. 2019; 177(2):478–491. https://doi.org/10.1016/j.cell.2019.01.048 PMID: 30929901

52. Xiang b, Wang Q, Lei W, Li M, Li Y, Zhao L, et al. Genes in immune pathways associated with abnormal white matter integrity in first-episode and treatment-naïve patients with schizophrenia. The British journal of psychiatry. 2019; 214(5):281–287. https://doi.org/10.1192/bjp.2018.297 PMID: 30722794

53. Melnik A, Tauber S, Dumrese C, Ullrich O, Wolf S. Murine adult neural progenitor cells alter their proliferative behavior and gene expression after the activation of Toll-like-receptor 3. European journal of microbiology & immunology. 2012; 2(3):239–248. https://doi.org/10.1556/EuJMI.2.2012.3.10 PMID: 24688771

54. Paridaen JTML, Janson E, Utami KH, Pereboom TC, Essers PB, Rooijen Cv, et al. The nucleolar GTP-binding proteins Gnl2 and nucleostemin are required for retinal neurogenesis in developing zebrafish. Developmental biology. 2011; 355(2):286–301. https://doi.org/10.1016/j.ydbio.2011.04.028 PMID: 21565180

55. Shrikumar A, Tian K, Shcherbina A, Avsec Ž, Banerjee A, Sharmin M, et al. TF-MoDISco v0. 4.4. 2-alpha. arXiv preprint arXiv:181100416. 2018;.

56. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome biology. 2007; 8(2):R24. https://doi.org/10.1186/gb-2007-8-2-r24 PMID: 17324271

57. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158(6):1431–1443. https://doi.org/10.1016/j.cell.2014.08.009 PMID: 25215497

58. Yutsudo N, Kamada T, Kajitani K, Nomaru H, Katogi A, Ohnishi YH, et al. fosB-null mice display impaired adult hippocampal neurogenesis and spontaneous epilepsy with depressive behavior. Neuropsychopharmacology. 2013; 38(5):895. https://doi.org/10.1038/npp.2012.260 PMID: 23303048

59. Velazquez FN, Caputto BL, Boussin FD. c-Fos importance for brain development. Aging (Albany NY). 2015; 7(12):1028. https://doi.org/10.18632/aging.100862

60. Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nature reviews Genetics. 2014; 15(4):234. https://doi.org/10.1038/nrg3663 PMID: 24614316

61. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. cell. 2006; 126(4):663–676. https://doi.org/10.1016/j.cell.2006.07.024 PMID: 16904174

62. Liao Y, Zhuang X, Huang X, Peng Y, Ma X, Huang ZX, et al. A bivalent securinine compound SN3-L6 induces neuronal differentiation via translational upregulation of neurogenic transcription factors. Frontiers in pharmacology. 2018; 9:290. https://doi.org/10.3389/fphar.2018.00290 PMID: 29674963

63. Paap RH, Oosterbroek S, Wagemans CM, von Oerthel L, Schellevis RD, Vastenhouw-van der Linden AJ, et al. FoxO6 affects Plxna4-mediated neuronal migration during mouse cortical development.

Proceedings of the National Academy of Sciences. 2016; 113(45):E7087–E7096. https://doi.org/10.1073/pnas.1609111113 PMID: 27791111

64. Sun Z, da Fontoura CS, Moreno M, Holton NE, Sweat M, Sweat Y, et al. FoxO6 regulates Hippo signaling and growth of the craniofacial complex. PLoS genetics. 2018; 14(10):e1007675. https://doi.org/10.1371/journal.pgen.1007675 PMID: 30286078

65. Liu W, Zhou H, Liu L, Zhao C, Deng Y, Chen L, et al. Disruption of neurogenesis and cortical development in transgenic mice misexpressing Olig2, a gene in the Down syndrome critical region. Neurobiology of disease. 2015; 77:106–116. https://doi.org/10.1016/j.nbd.2015.02.021 PMID: 25747816

66. Genestine M, Lin L, Durens M, Yan Y, Jiang Y, Prem S, et al. Engrailed-2 (En2) deletion produces multiple neurodevelopmental defects in monoamine systems, forebrain structures and neurogenesis and behavior. Human molecular genetics. 2015; 24(20):5805–5827. https://doi.org/10.1093/hmg/ddv301 PMID: 26220976

67. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

68. Yan J, Qiu Y, Dos Santos AMR, Yin Y, Li YE, Vinckier N, et al. Systematic analysis of binding of transcription factors to noncoding variants. Nature. 2021; 591(7848):147–151. https://doi.org/10.1038/s41586-021-03211-0 PMID: 33505025

69. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic acids research. 2016; 44(11):e107–e107. https://doi.org/10.1093/nar/gkw226 PMID: 27084946

70. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. The American Journal of Human Genetics. 2016; 98(6):1114–1129. https://doi.org/10.1016/j.ajhg.2016.03.029 PMID: 27236919

71. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics. 2014; 94(4):559–573. https://doi.org/10.1016/j.ajhg.2014.03.004 PMID: 24702953

72. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nature biotechnology. 2010; 28(10):1045. https://doi.org/10.1038/nbt1010-1045 PMID: 20944595

73. Avsec Z, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, et al. Kipoi: accelerating the community exchange and reuse of predictive models for genomics. BioRxiv. 2018; p. 375345.

74. Trevino AE, Sinnott-Armstrong N, Andersen J, Yoon SJ, Huber N, Pritchard JK, et al. Chromatin accessibility dynamics in a model of human forebrain development. 2020; 367 (6476).

75. Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, et al. Evolutionary changes in promoter and enhancer activity during human corticogenesis. 2015; 347(6226):1155–1159.

76. Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru MD, Reach SM, et al. An atlas of chromatin accessibility in the adult human brain. Genome research. 2018; 28(8):1243–1252. https://doi.org/10.1101/gr.232488.117 PMID: 29945882

77. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. 2018; 362 (6420).

78. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10); 2010. p. 807–814.

79. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

80. Nesterov YE. A method for solving the convex programming problem with convergence rate O (1/k^ 2). In: Dokl. akad. nauk Sssr. vol. 269; 1983. p. 543–547.

81. Wang k, Li M, Hakon H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic acids research. 2010; 38(16):e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685

82. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020; 581(7809):434–443. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654

83. Consortium GP, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68. https://doi.org/10.1038/nature15393

84. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS computational biology. 2014; 10(7):e1003711. https://doi.org/10.1371/journal.pcbi.1003711 PMID: 25033408

85. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics. 2016; 32(14):2205–2207. https://doi.org/10.1093/bioinformatics/btw203 PMID: 27153639

86. Lee D. LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics. 2016; 32(14):2196–2198. https://doi.org/10.1093/bioinformatics/btw142 PMID: 27153584

87. Shigaki D, Adato O, Adhikari AN, Dong S, Hawkins-Hooker A, Inoue F, et al. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. Human mutation. 2019; 40(9):1280–1291. https://doi.org/10.1002/humu.23797 PMID: 31106481

88. Koo PK, Eddy SR. Representation Learning of Genomic Sequence Motifs with Convolutional Neural Networks. BioRxiv. 2018; p. 362756.

89. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org; 2017. p. 3145–3153.

90. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034. 2013;.