Brain-ResNet

This manuscript (permalink) was automatically generated from sq-96/Brain-ResNet@c54eabc on June 4, 2020.

Authors

- **Sheng Qian**

Abstract

Decoding the regulatory behavior of DNA sequences and the functional effects of noncoding variants is a preeminent challenge in understanding the mechanisms of gene regulation. This is also important for the genetics of common diseases, as most disease-associated variants are located in noncoding regions of the genome. Recently, Convolutional Neural Networks (CNNs) based methods have been developed to predict genome-wide chromatin profiles in various cellular contexts. However, these tools and resources were often trained in cell lines or bulk tissues that are not necessarily disease-related. This is particularly an issue for neuropsychiatric disorders, where the most relevant cell and tissue types are missing in the training data used by current tools.

Introduction

Next-generation sequencing(NGS) technologies have given rise to the development of many sequencing assays such as ATAC-seq[1], DNase-seq[2], ChIPseq, RNA-seq, and FIAR-seq that measure the epigenomic landscapes across many cellular contexts, including histone marks, TF binding and chromatin accessibility. These epigenomic annotations aid the characterization of noncoding genomic variants and show promises in assessing disease-associated variants and understanding the underlying transcription machinery. There has been a joint effort to survey the noncoding part of the human genome by the community, and numerous noncoding genomic sites have been statistically identified for association with complex traits. Leveraging these resources, researchers have developed machine learning models to learn features of DNA sequences that predict chromatin profiles such as protein binding sites, chromatin accessibility, histone marks and methylation of DNA sequences. Once a sequence based model is trained to predict a certain epigenomic feature, a researcher can use it to predict the likely epigenomic effect of a DNA variant.

Results

1. Enrichment of ASoC Variants

To validate our predicion model, we first performed enrichment analysis of allele-specific open-chromatin (ASoC) variants. Genetic variants prioritized by our prediction model are expected to have large functional effects. We hypothesize that our predictions are enriched for genetic variants with some known functions. ASoC variants have been established to be functional in brain, impacting gene expresison, histone modification and DNA methylation[3]. We obtained ASoC variants in neural progenitor cells (NPC) and glutamatergic (iN-Glut) neurons from a neuron ATAC-Seq study[3]. We then acquired all single nucleotide variants in open chromatin regions of NPC and iN-Glut and prioritized them by our NPC and iN-Glut Brain-ResNet scores. The top 10,000 predicted genetic variants show 4 fold enrichment of ASoC variants in NPC and iN-Glut. To show the strength of our model, we also prioritized genetic variants within open chromatin regions by Functional significance (Funsig) score and CADD score[4,5]. Funsig is a measure of the signficance of magnitude of predicted chromatin effect and evolutionary conservation, and CADD score is a measure of the deleteriousness of genetic variants. As shown in Fig1, our Brain-ResNet scoring significantly outperforms Funsig and CADD scoring. This gaining may arise from two apsects. First, our model uses functional genomic data from matched cell types, which could more accurately reveal the chromatin status. Second, our model uses ResNet architecture and is based on transfer learning, which could more precisely learn regulatory codes from DNA sequences. To further address the importance of matched cell types, we used Brain-ResNet scores from the other 30 cell types to prioritize genetic variants in NPC and iN-Glut. As shown in Fig2, top predictions prioritized by matched cell types generally have higher enrichment of ASoC variants.

2. Sign Consistentcy

Functional genetic variants either increase or decrease intensity of a certein activity in the genome. To test if our model can precisely predict the effect size and the direction of effect, we applied our prediction model to NPC and iN-Glut ASoC variants and compared the observed allelic imbalance and the predicted difference in functional effects between reference and alternative alleles. As shown in Fig3, Our prediction model tracks the observed allelic imbalance ratio with a correlation of 0.44 and 0.40. Notably, we found 70% variants show consistent sign in observed allelic imbalace and estimated effect, which demonstrates that the prediction model accurately captures the direction of effect.

3. Evolutionary Constraint

Evolutionary constraint has shown to be useful in identifying functionally important regions[6]. Leveraging this strategy, we calculated GERP score for top predicted variants and randomly sampled variants in 31 cell types. GERP score measures the number of substitutions “rejected” by evolutionary constraint and higehr GERP score indicates greater magnitude of evolutionary constraint[7]. As shown in Fig4, for most cell types, our prediction model successfully prioritized genetic variants that are under higher evolutionary constraint and are more likely to have actual biological functions.

4. Purifying Selection

Because DNA variations are more likely to be deleterious than beneficial, negative selection are required to remove damaing mutations and maintain the stability of biology system [8]. This is especially true for functionally important variants, whose change may disrupt essential biological functions. To investigate if our Brain-ResNet score could indicate functional effects, we obtained minor allele frequency from gnomAD database for all variants within peak regions of 31 chromatin profiles and plotted them against their predicted functional effects. As shown in Fig5, there is a clear negtive correlation between minor allele frequency and Brain-ResNet score. Genetic variants with larger predicted functional effects tend to have lower minor allele frequency, which indicates the acting of negtive selection. This evidence suggests that our Brain-ResNet score is a good predictor of functional importance.

Figures

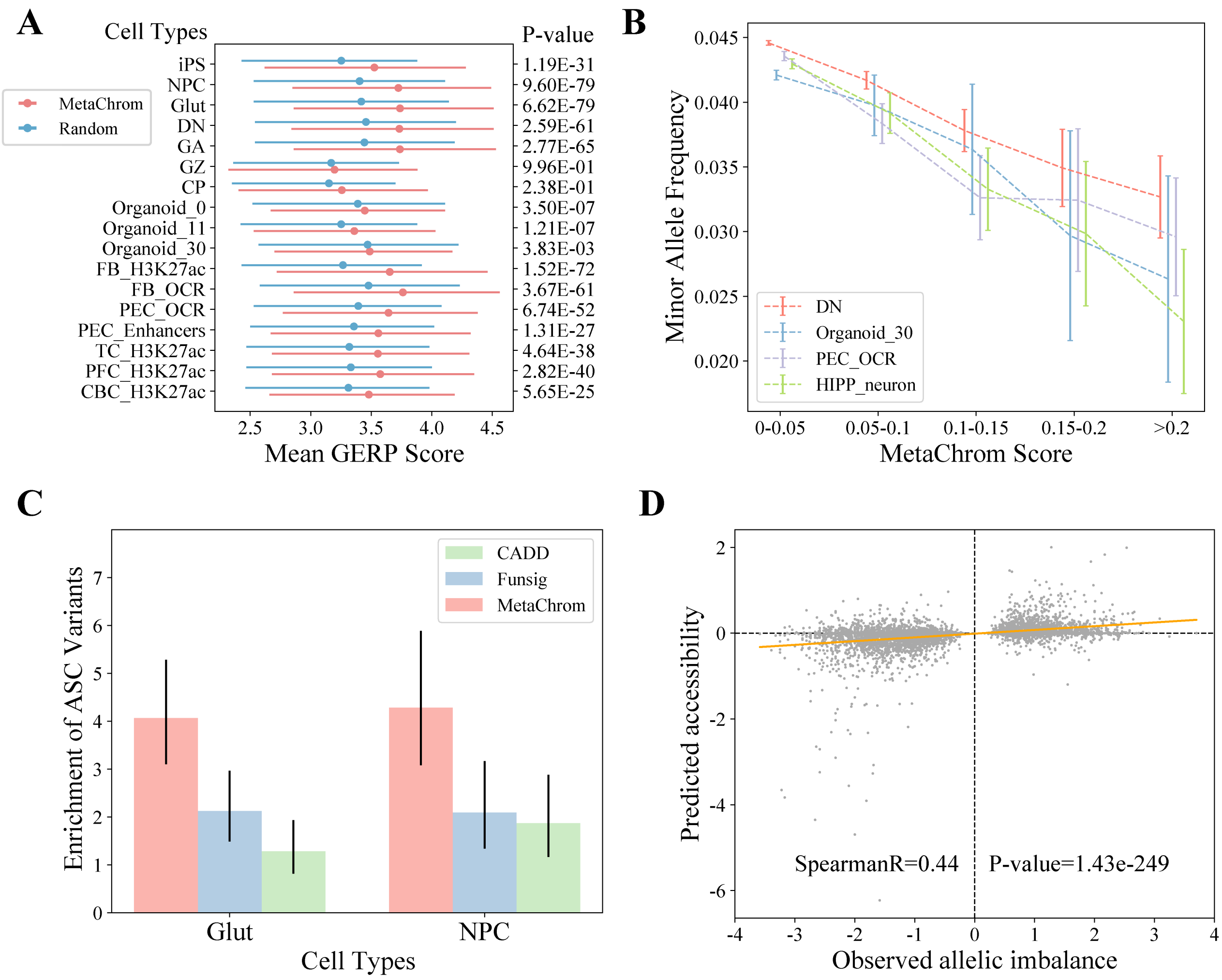

Figure 1: Validation of MetaChrom predicted functional variants. (A) Distribution of GERP scores between MetaChrom predicted functional variants and random variants in fetal brain cell types. (B) Minor allele frequency by MetaChrom score for variants within open chromatin region in 4 cell types. (C) Enrichment of ASC variants for predicted functional variants identified by MetaChrom, Funsig and CADD score in Glut and NPC cells. (D) Scatter plot comparing the observed allelic imbalace and MetaCHrom predicted effect on chromatin accessibility of ASC variants.

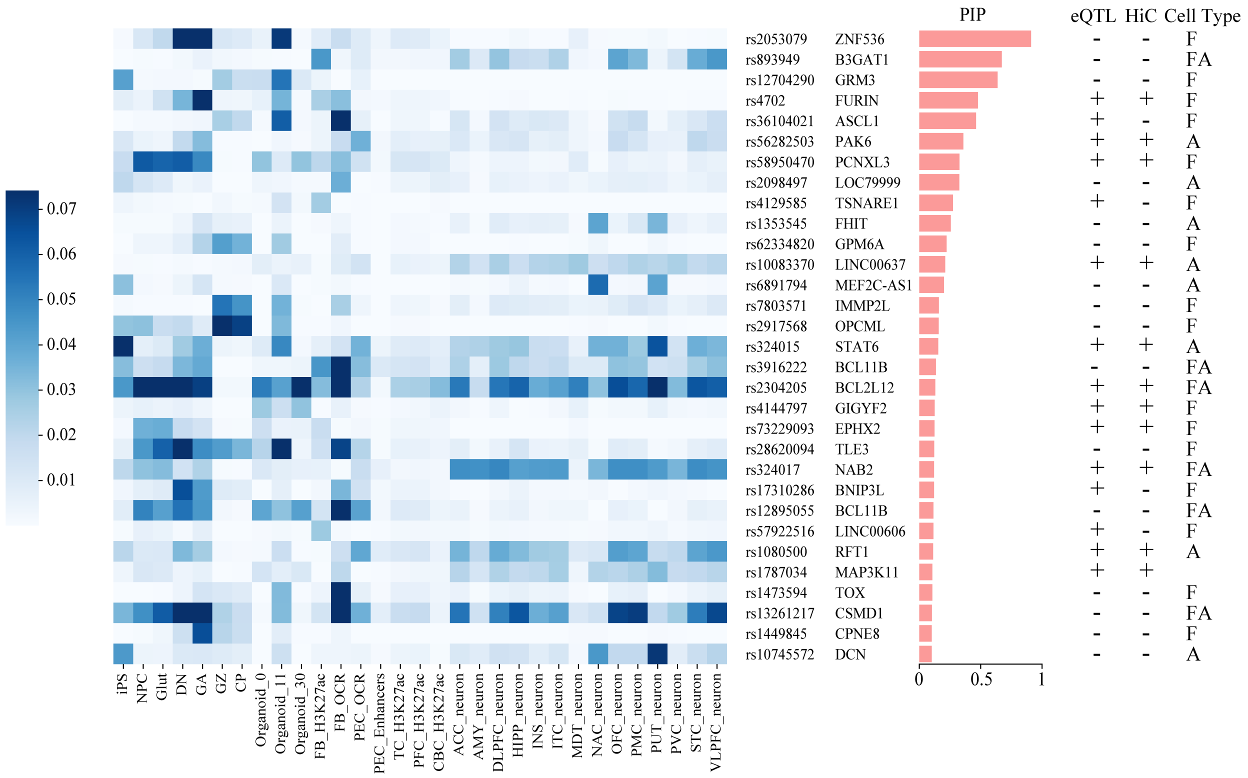

Figure 2: **Heatmap.** Heatmap showing functional effects of credible set SNPs in 31 cell types.

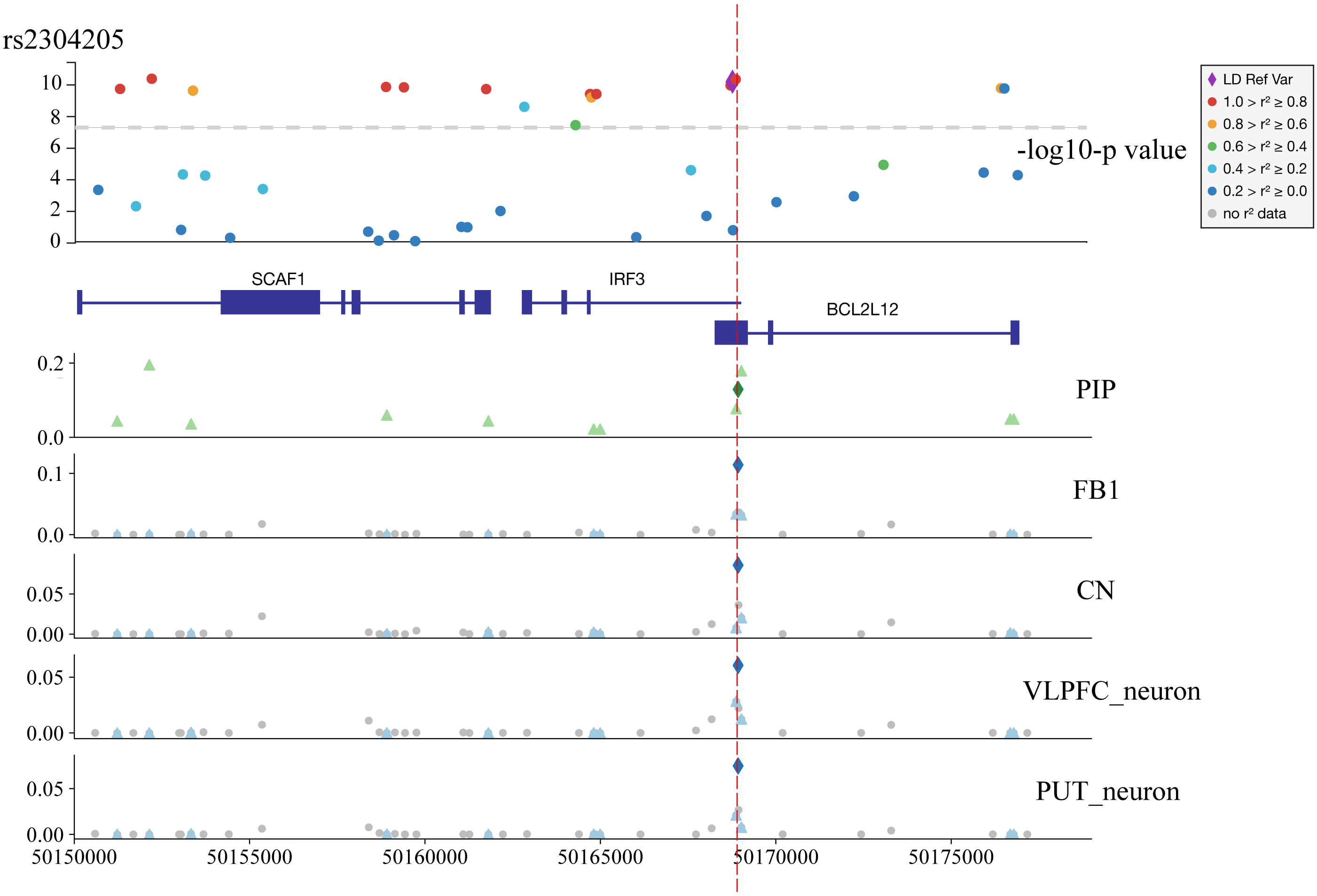

Figure 3: **Tracks.** Scatter plot showing pvalue, pip and functional effects of the candidate SNP.

1. Enrichment of ASC Variants

All single nucleotide variants (SNVs) within functional regions (open chromatin or H3K27ac) in each cell type are retrieved from 1000 Genomes Project. We calculate MetaChrom score, Funsig score and CADD score for all SNVs. Funsig score is obtained from the DeepSEA Server and CADD score is obtained from annovar. We define the top 10,000 variants ranked by MetaChrom score, Funsig score and CADD score in descending order as MetaChrom, Funsig and CADD predicted functional variants, respectively. ASC variants in NPC and Glut cells are obtained from a neuron ATAC-seq paper. In the neuron ATAC-seq study, iPSCs of 20 individuals are first differentiated into neural progenitor cells (NPC) and gutamatergic (iN-Glut) neurons. Then, ATAC-seq is performed, and 5,611 and 3,547 ASoC SNPs are identified in NPC and iN-Glut cells by allelic imbalance test, respectively. We count the number of ASC variants in predicted functional variants and control variants, and the Enrichment of ASC variants is calculated by fisher exact test.

2. Sign Consistentcy

For ASC variants, we define the observed allelic imbalance as log(ref reads/alt reads) and the predicted effect on chromatin accessibility as log(ref pred/alt pred). Correlation between observed allelic imbalance and predicted effect on chromatin accessibility is calcualted by Spearman’s rank correlation coefficient.

3. Evolutionary Constraint

GERP scores of MetaChrom predicted functional variants and control variants are obtained from Annovar. P value is calculated by the Wilcoxon Rank-Sum Test.

4. Purifying Selection

Minor allele frequency of MetaChrom predicted functional variants and control variants is obtained from gnomAD. The MetaChrom score ranges from 0 to 1. We splitted variants into 5 bins according to the MetaChrom score, namely 0-0.05, 0.05-0.1, 0.1-0.15, 0.15-0.2 and 0.2-1.0. In each bin, the mean of minor allele frequency and standard error of the mean are calcualated.

References

1. **ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide**
Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, William J. Greenleaf
Current Protocols in Molecular Biology (2015-01-05) https://doi.org/gdwsxx
DOI: 10.1002/0471142727.mb2129s109 · PMID: 25559105 · PMCID: PMC4374986

2. **DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells**
L. Song, G. E. Crawford
Cold Spring Harbor Protocols (2010-02-01) https://doi.org/d7rhg8
DOI: 10.1101/pdb.prot5384 · PMID: 20150147 · PMCID: PMC3627383

3. **Allele-specific open chromatin in human iPSC neurons elucidates functional non-coding disease variants**
Siwei Zhang, Hanwen Zhang, Min Qiao, Yifan Zhou, Siming Zhao, Alena Kozlova, Jianxin Shi, Alan R. Sanders, Gao Wang, Subhajit Sengupta, … Jubao Duan
bioRxiv (2019-11-01) https://doi.org/ggtqw2
DOI: 10.1101/827048

4. **Predicting effects of noncoding variants with deep learning–based sequence model**
Jian Zhou, Olga G Troyanskaya
Nature Methods (2015-08-24) https://doi.org/gcgk8g
DOI: 10.1038/nmeth.3547 · PMID: 26301843 · PMCID: PMC4768299

5. **A general framework for estimating the relative pathogenicity of human genetic variants**
Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, Jay Shendure
Nature Genetics (2014-02-02) https://doi.org/f5s57j
DOI: 10.1038/ng.2892 · PMID: 24487276 · PMCID: PMC3992975

6. **Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes**
D. L. Goode, G. M. Cooper, J. Schmutz, M. Dickson, E. Gonzales, M. Tsai, K. Karra, E. Davydov, S. Batzoglou, R. M. Myers, A. Sidow
Genome Research (2010-01-12) https://doi.org/b9qjng
DOI: 10.1101/gr.102210.109 · PMID: 20067941 · PMCID: PMC2840986

7. **Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++**
Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, Serafim Batzoglou
PLoS Computational Biology (2010-12-02) https://doi.org/csr7f4
DOI: 10.1371/journal.pcbi.1001025 · PMID: 21152010 · PMCID: PMC2996323

8. **Negative Selection | Learn Science at Scitable**https://www.nature.com/scitable/topicpage/negative-selection-1136/