Brain-ResNet

This manuscript (permalink) was automatically generated from sq-96/Brain-ResNet@b6b195d on April 12, 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 lagre functional effects. One way to test our hypothesis is calculating the enrichment of some genetic variants with konwn functional effects in our top predictions. ASoC variants are overrepresented in brain enhancers, transcription-factor-binding sites, and quantitative-trait-loci associated with gene expression, histone modification, and DNA methylation. We obtained ASoC variants in neural progenitor cells (NPC) and glutamatergic (iN-Glut) neurons from a neuron ATAC-seq study. We then acquired single nucleotide variants in open chromatin regions of NPC and iN-Glut by mapping against 1000 Genome and prioritized them by our Brain-ResNet scores. The top 10,000 predicted genetic variants show 4 fold enrichment of ASoC variants in NPC and iN-Glut. AS a comparsion, we also prioritized genetic variants with Functional significance (Funsig) score and CADD score. 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 outperforms Funsig and CADD scoring in terms of identifying function variants.

2. Sign Consistentcy

Next, 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. 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 taht the prediction model accurately captures the direction of effect.

3. Evolutionary Constraint

Functionally important variants tend to be under evolutionary constraint. We calculated GERP score, which measures the number of substitutions “rejected” by evolutionary constraint, for top predicted variants and randomly sampled variants in 31 cell types. Higehr GERP score indicates greater magnitude of evolutionary constraint. As shown in fig3, for most cell types, our prediction model successfully prioritized genetic variants that are under higher evolutionary constraint and more likely to have siginificant biological functions.

4. Purifying Selection

Because more DNA changes are harmful than are beneficial, negative selection plays an important role in maintaining the long-term stability of biological structures by removing deleterious mutations. This is especially true for functionally important variants, whose change may disrupt essential biological functions. 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 fig4, variants with larger predicted functional effects tend to have lower minor allele frequency, which indicates the acting of purifying selection.

Figures

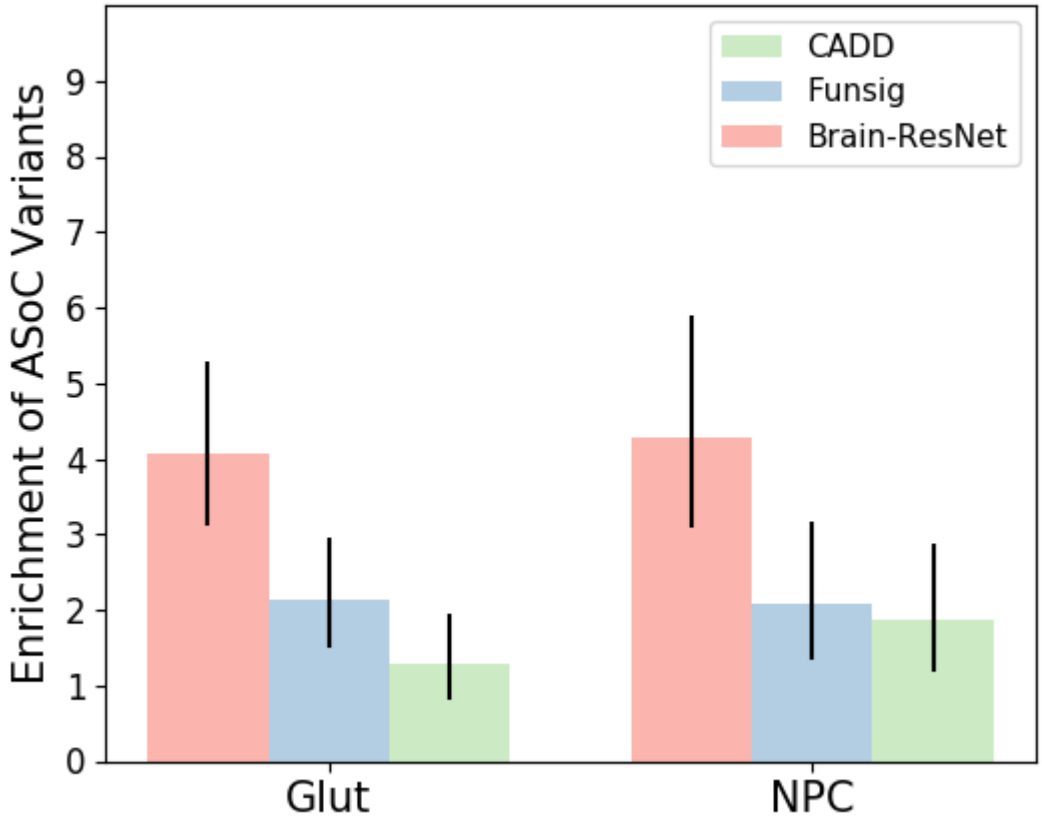


Figure 1: ASoC Enrichment. Bar plot comparing the enrichment of allele specific open chromatin variants among three groups in two cell types.

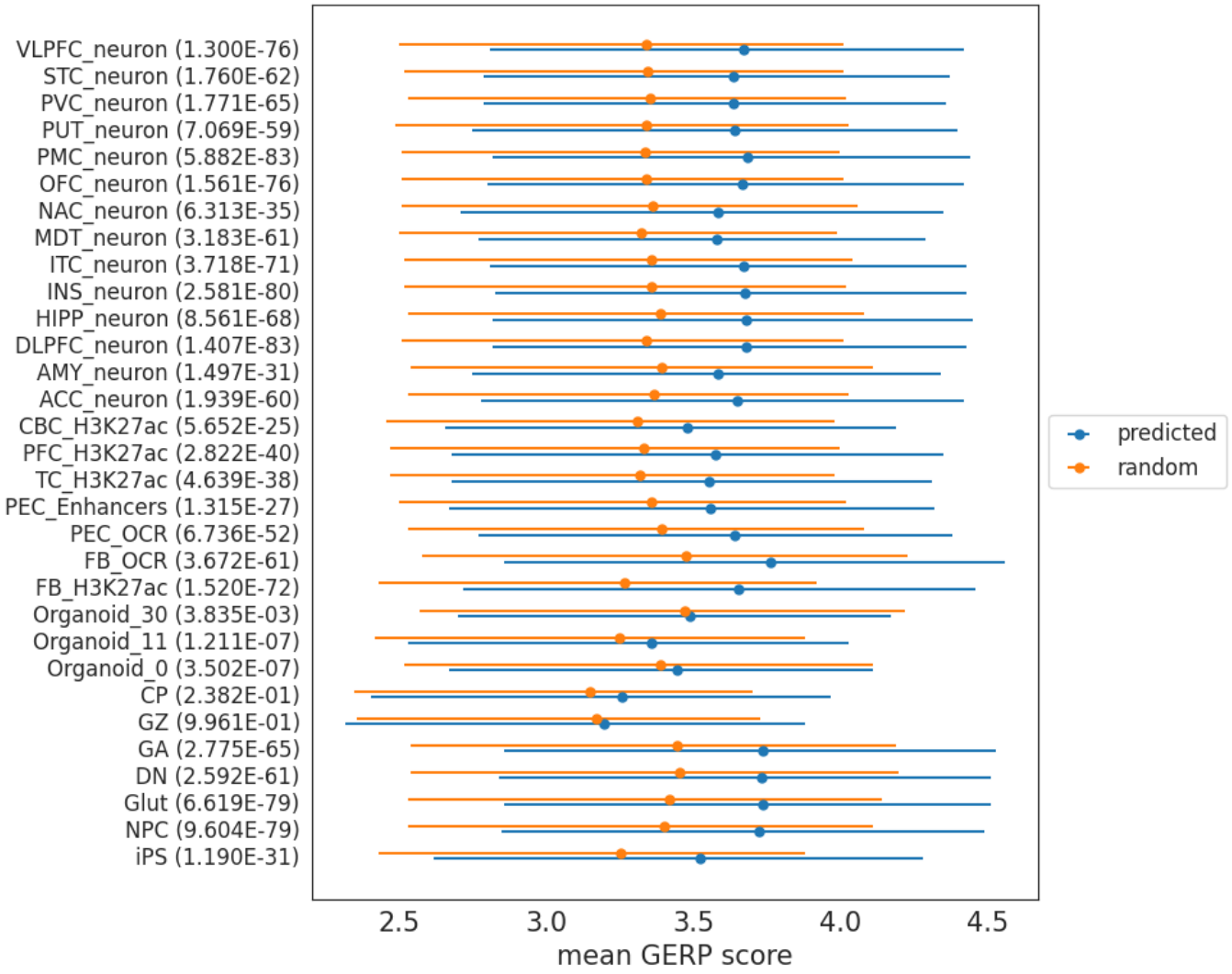


Figure 2: GERP Score Distribution. Bar plot comparing the evolutionary constraint between two groups in 31 cell types.

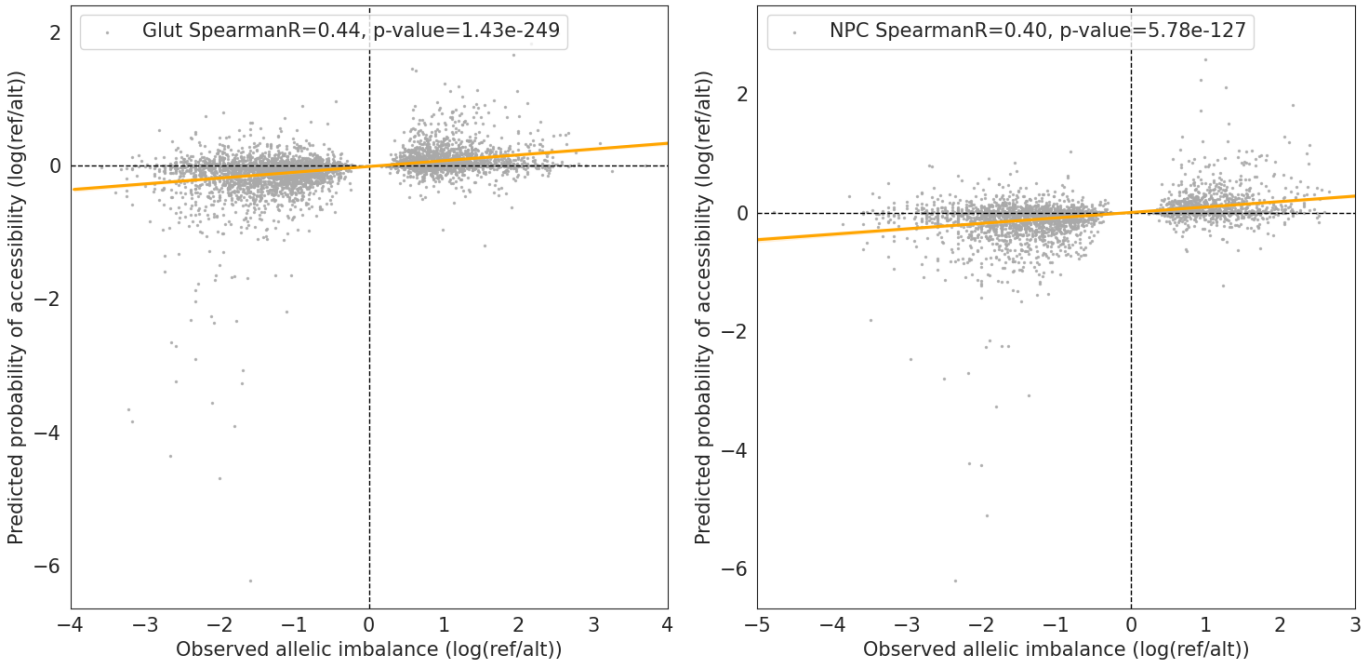
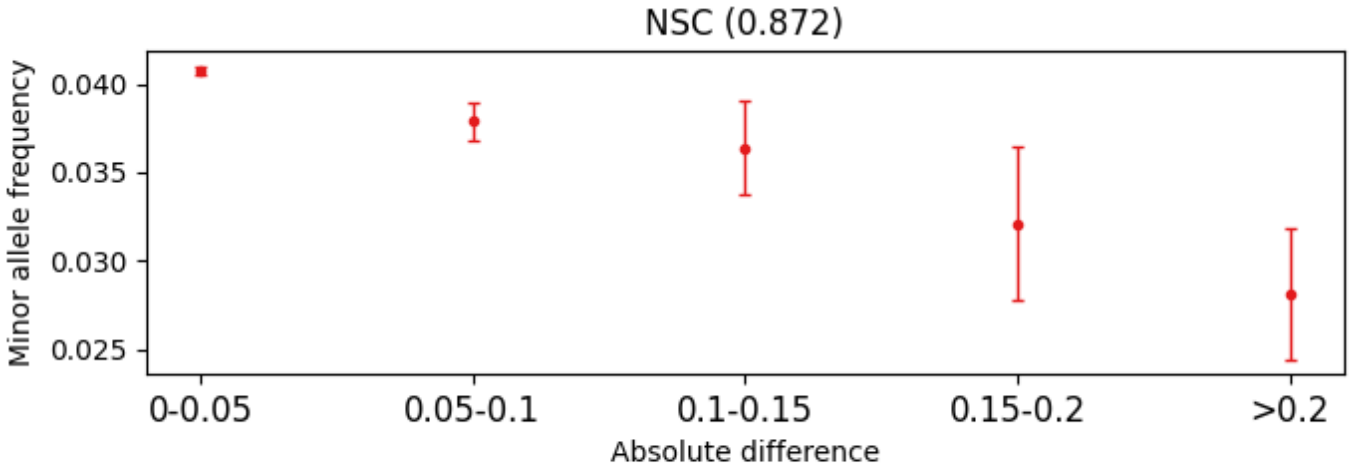
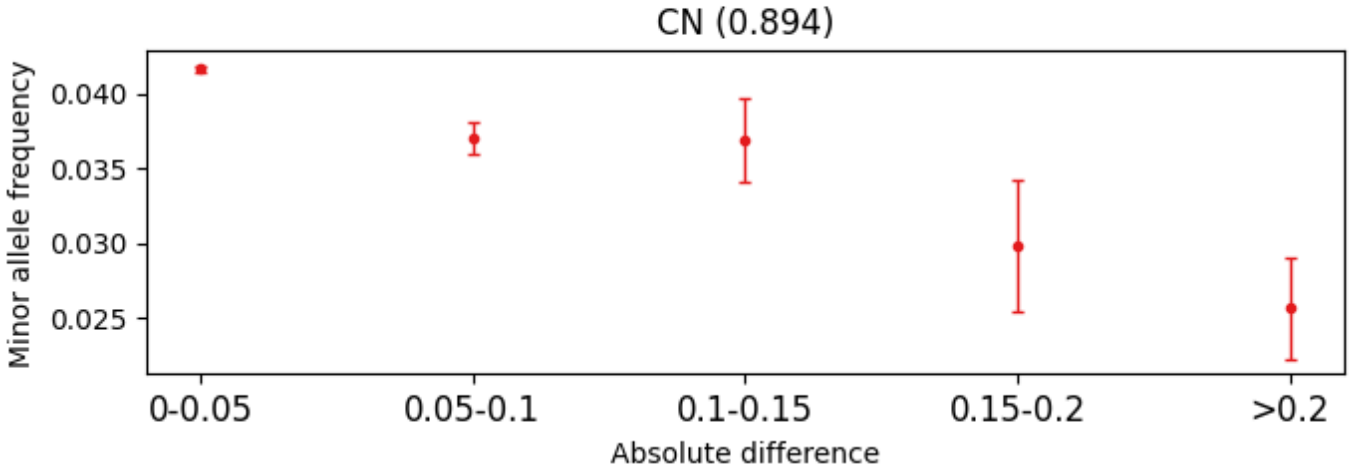
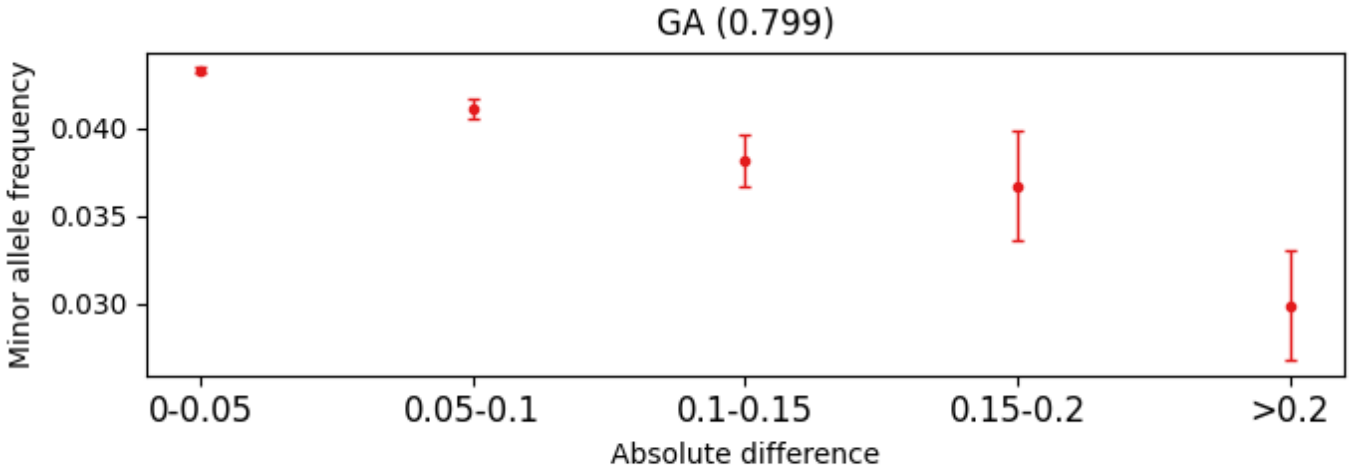


Figure 3: Sign Consistency. Scatter plot comparing the observed allelic imbalance and the predicted difference in functional effects between reference and alternative alleles.


NSC (0.872)
Minor allele frequency
0.040
0.035
0.030
0.025
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
>0.2
Absolute difference


CN (0.894)
Minor allele frequency
0.040
0.035
0.030
0.025
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
>0.2
Absolute difference


GA (0.799)
Minor allele frequency
0.040
0.035
0.030
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
>0.2
Absolute difference

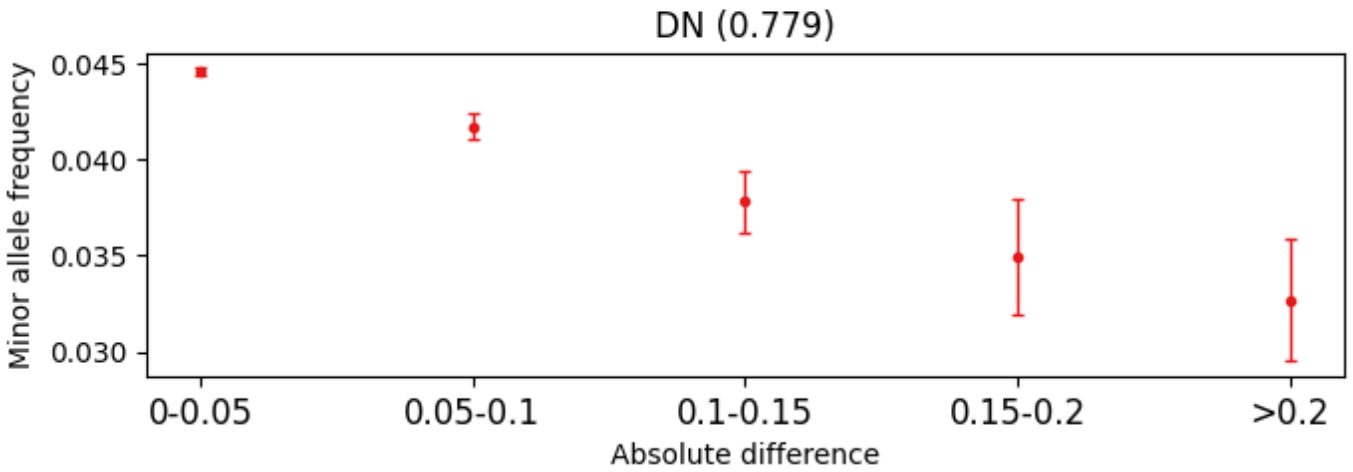

DN (0.779)
Minor allele frequency
0.045
0.040
0.035
0.030
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
>0.2
Absolute difference

Figure 4: Minor Allele Frequency. Scatter plot showing the negtive correlation between minor allele frequency and Brain-ResNet predicted scores.

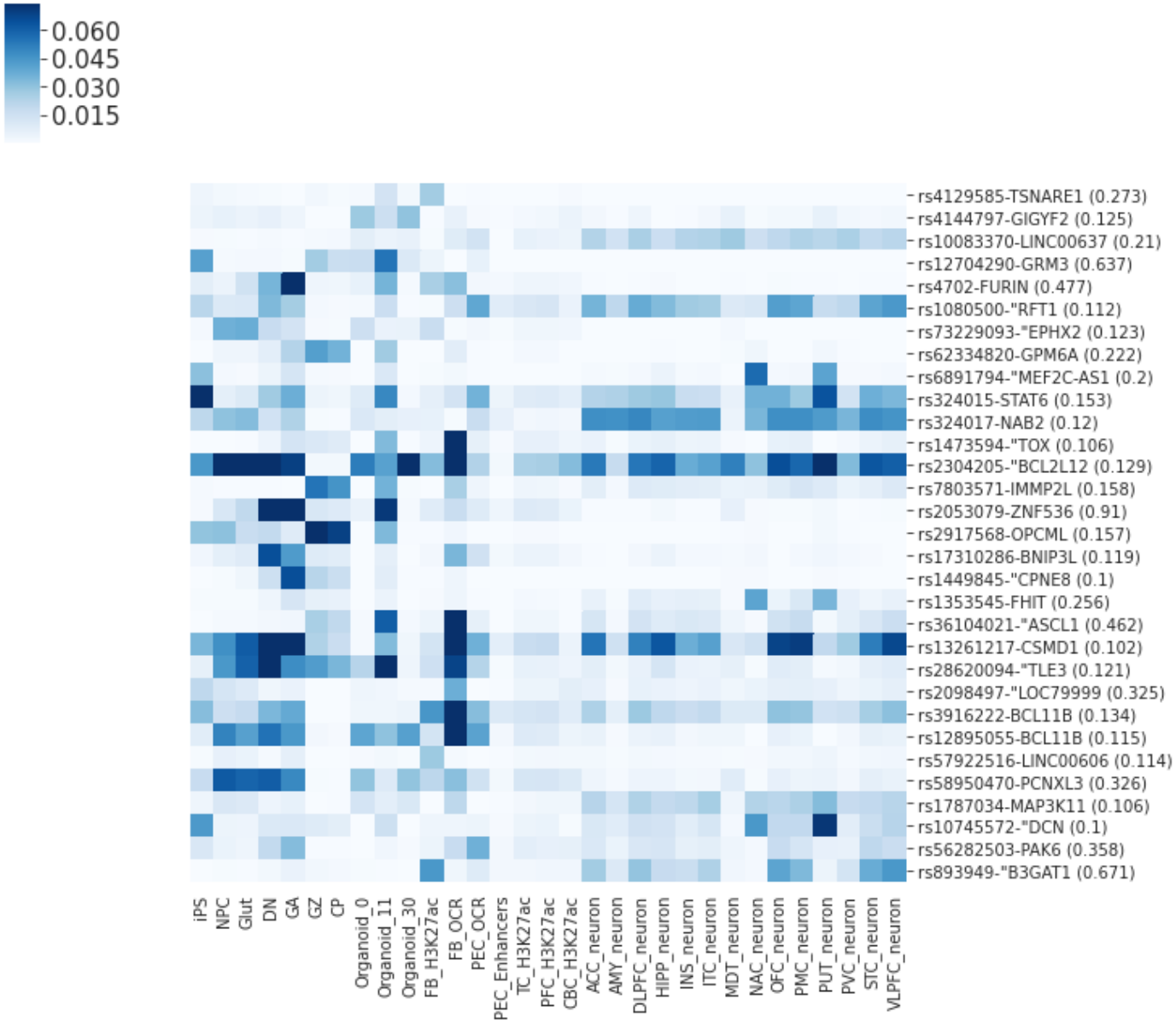


Figure 5: Heatmap. Heatmap showing functional effects of credible set SNPs in 31 cell types.

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