Brain-ResNet

This manuscript (permalink) was automatically generated from sq-96/Brain-ResNet@b054afa 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

2. Evolutionary Constraint

3. Sign Consistentcy

4. Negtive Selection

Figures

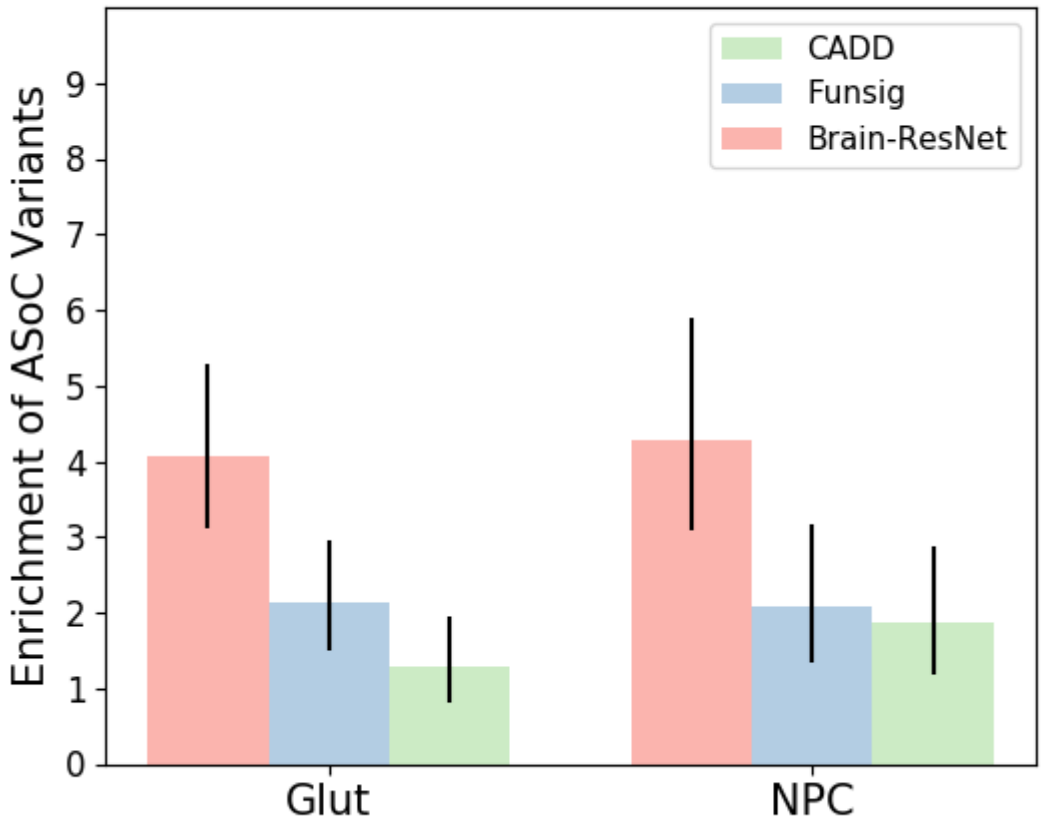


Figure 1: ASoC Enrichment. Bar plot comparing the enrichment of allele specific open chromatin variants among three groups in two cell types.

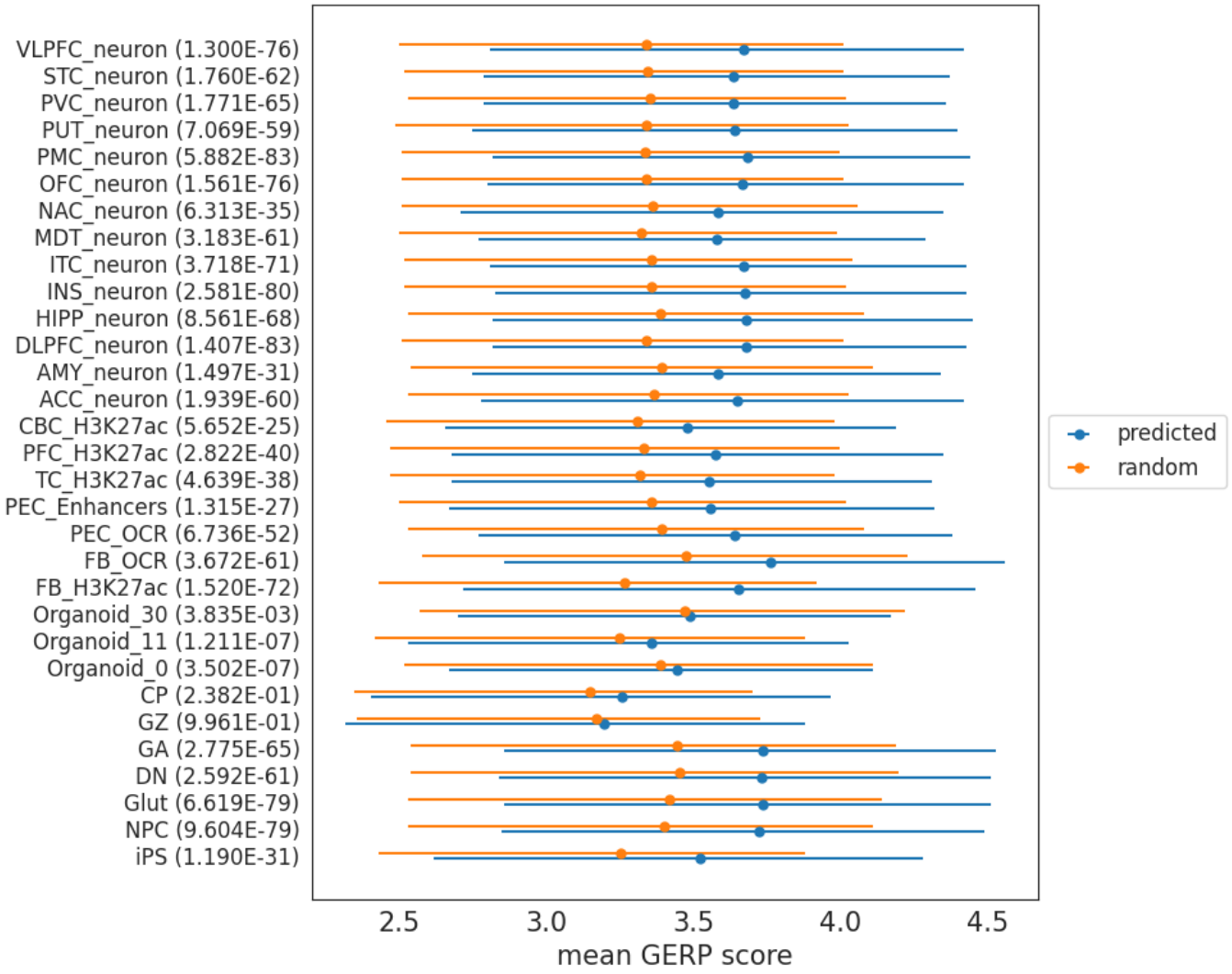


Figure 2: GERP Score Distribution. Bar plot comparing the evolutionary constraint between two groups in 31 cell types.

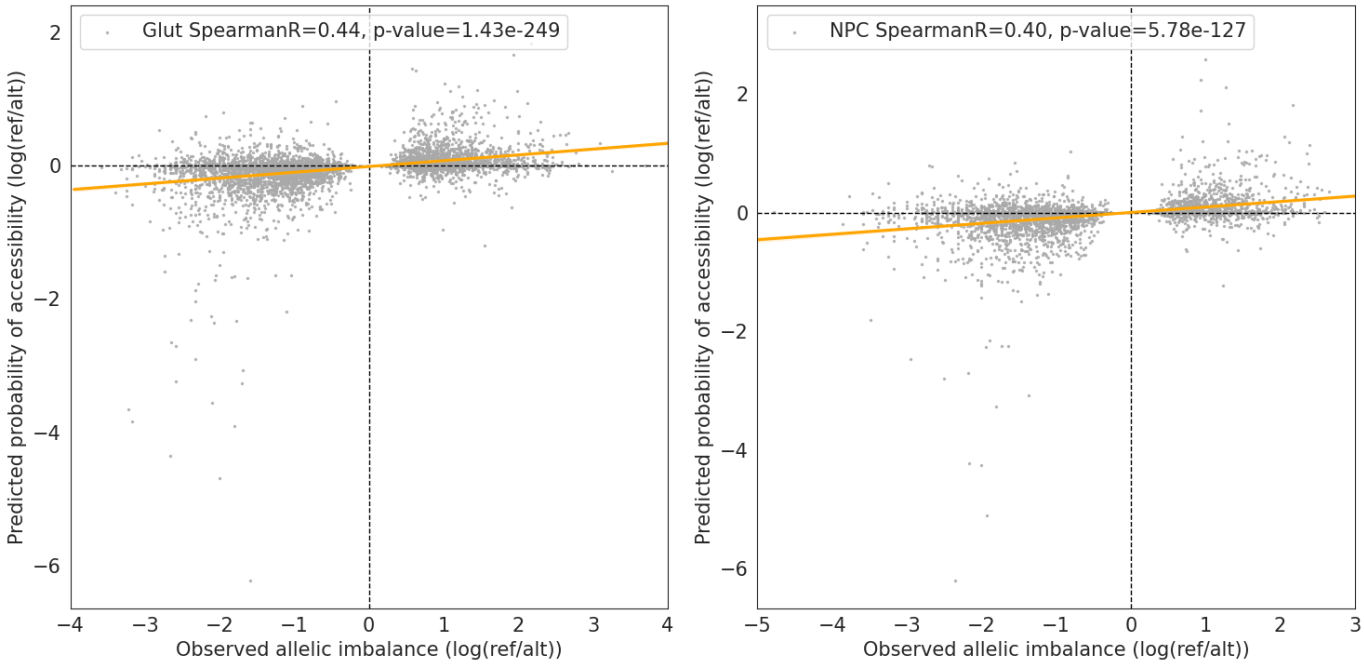
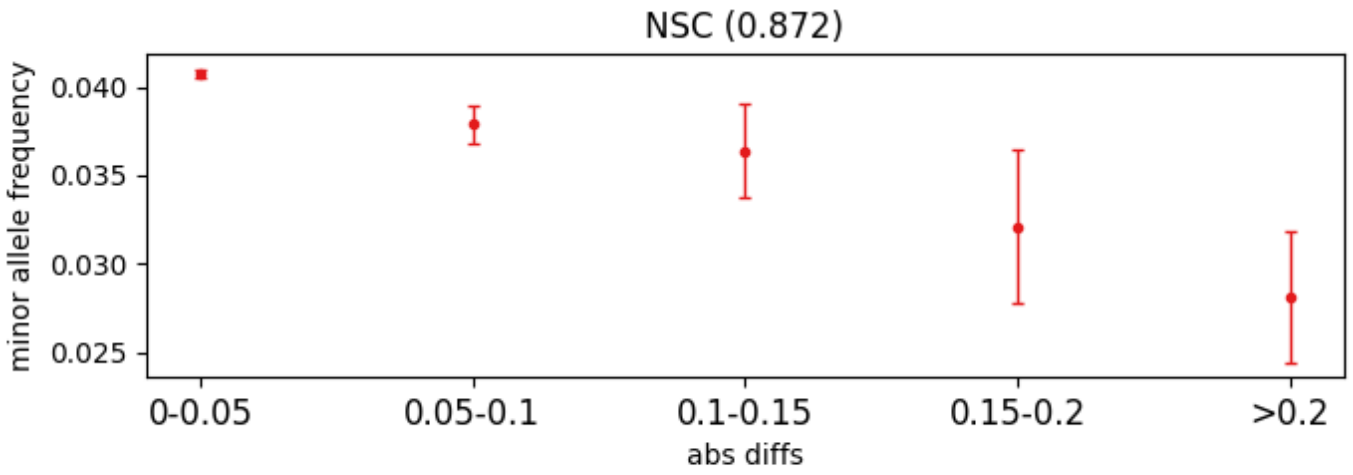
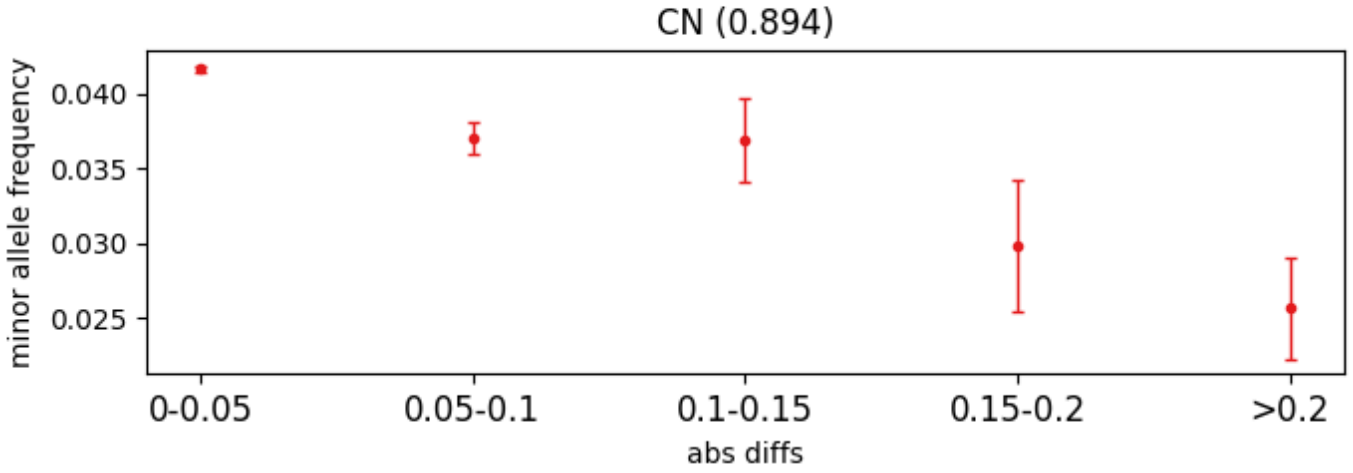
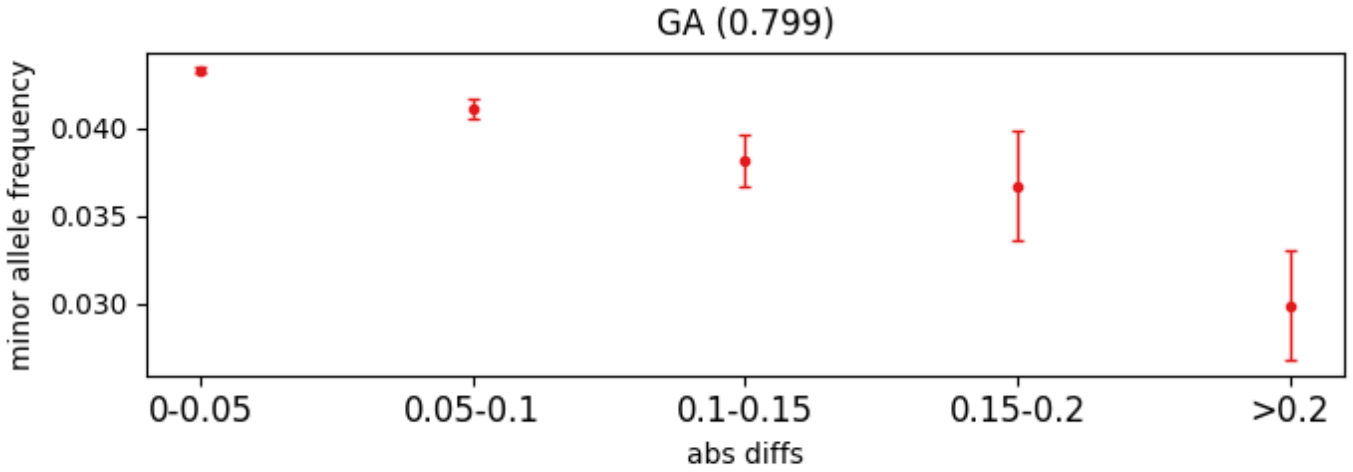


Figure 3: Sign Consistency. Scatter plot comparing the observed allelic imbalance and the predicted difference in chromatin activity between reference and alternative alleles.


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
abs diffs


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
abs diffs


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
abs diffs

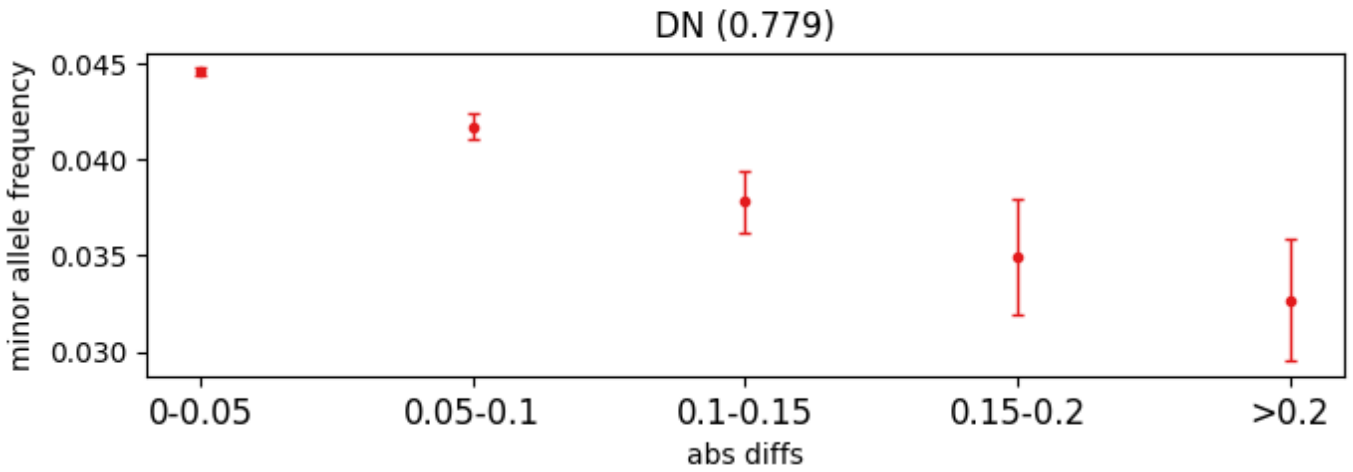

DN (0.779)
minor allele frequency
0.045
0.040
0.035
0.030
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
>0.2
abs diffs

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

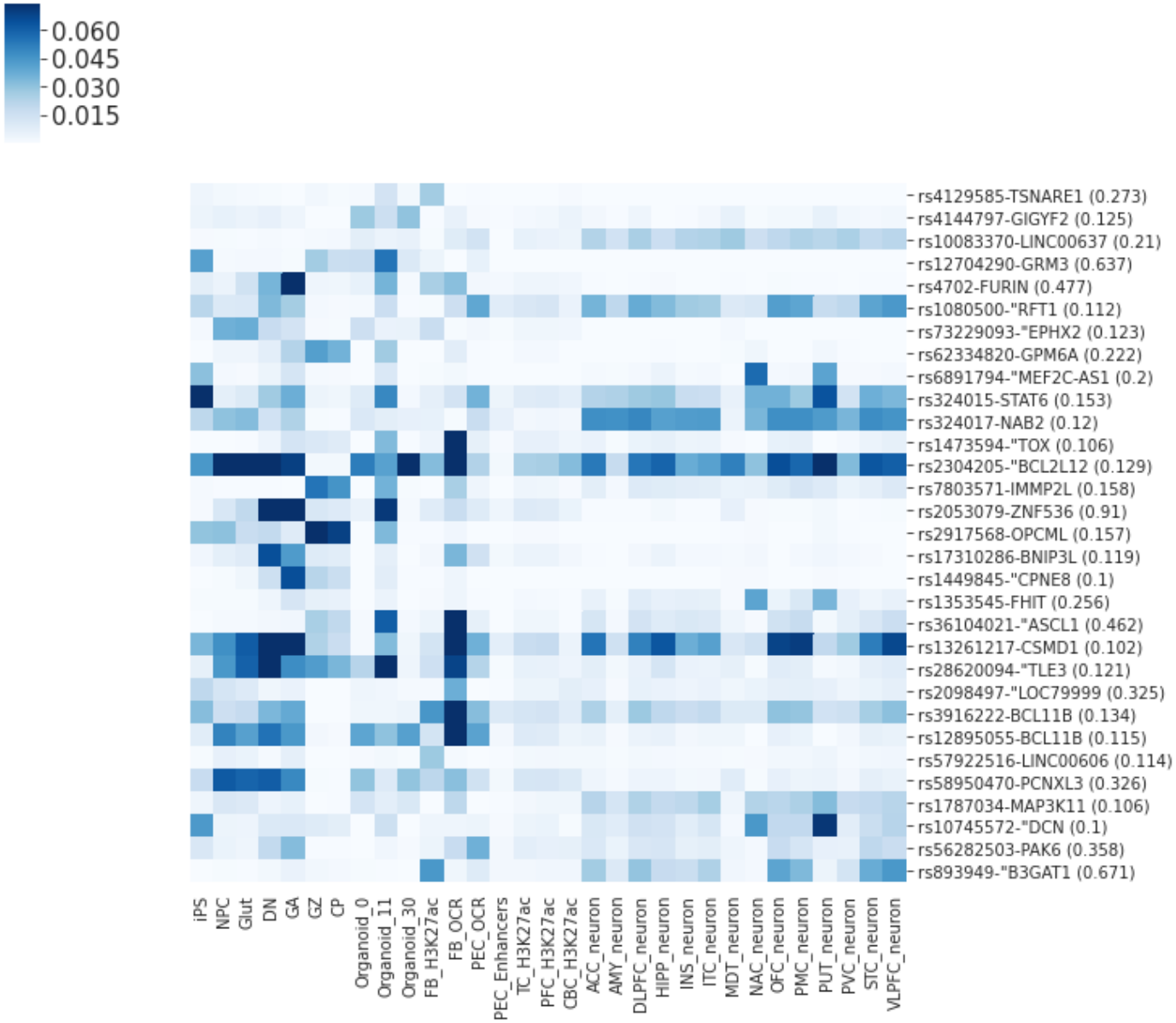


Figure 5: Heatmap. Heatmap showing functional effects of credible set SNPs in 31 cell types.

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