Submitter Name: Andy Chen Submitted Email: andychen@iu.edu
PI Name: Yunlong Liu PI email: yunliu@iu.edu

Functional Screening of Regulatory Variants Combined with Genome-wide Association and Machine Learning Identifies Causal Regulatory Mechanisms Impacting Substance Use Disorders

Andy B. Chen[1,2], Xuhong Yu[1], Xiaona Chu[1], Hongyu Gao[1,2,4], Jill L. Reiter[1,3], Xiaoling Xuei[1,4], Dongbing Lai[1], Yue Wang[1], Howard J. Edenberg[1,3], Yunlong Liu[1,2,]

[1]Department of Medical & Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA; [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA; [3]Department of Biochemistry & Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA; [4]Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN, USA

Non-coding regulatory elements such as in the 3' untranslated regions (3'-UTR) and enhancer regions can regulate gene expression. Variants in these regions are implicated in many diseases, but their mechanisms are less straightforward than those in the coding sequence, which may directly alter structures of proteins.

Using a massively parallel reporter assay, we can evaluate the functional effect of thousands of such variants by inserting their sequences into a reporter plasmid and observing the resulting gene expression changes in a transfected cell line. However, despite the high-throughput nature of these assays, the number of possible variants to be evaluated is much larger than could feasibly be performed.

To address this, we built a machine learning model to predict the potential outcomes of the MPRAs. Our multi-task model consisted of a convolutional neural network layer (to model motif-like sequences) and a long short-term memory layer (to model interactions between regulatory elements) trained to use the reference and alternative sequences evaluated by the MPRA to predict both sequence activity and variant impact. Using the results of MPRA experiments testing 9,550 3'-UTR variants (918 significant) and 23,122 enhancer variants (4,456 significant), we trained models to predict the impact of novel variants.

Using these predictions, we leverage a larger pool of regulatory variants to integrate impact with genome-wide association to identify genes that contribute to substance use disorders. This approach has uncovered potential new molecular mechanisms of addiction and potential therapeutic targets, showcasing the power of integrating diverse genetic and computational approaches.