A Genome-wide Association Analysis with Cocaine Use Disorder Accounting for both Phenotypic Heterogeneity and Gene-Environment Interplay

Jiangwen Sun[1], Henry R Kranzler[2], Joel Gelernter[3], Jinbo Bi[1]

[1] Department of Computer Science and Engineering, University of Connecticut; [2] Center for Studies of Addiction, University of Pennsylvania Perelman School of Medicine; [3] Department of Psychiatry, Yale University

To identify novel genetic variants that are associated with cocaine use disorder, we first performed cluster analysis to derive traits with reduced phenotypic heterogeneity. Then, we conducted a genome-wide association study (GWAS) using a Gaussian Mixture model that detects genetic associations when considering modulatory effects from environmental variables. The cluster analysis of a study sample (N=9965) identified five relatively homogeneous subgroups, two of which, i.e., Subtypes 4 (N=3,258) and 5 (N=1,916), comprised heavy cocaine users and had high heritability estimates (h2=0.66 and 0.64, respectively). Using logistic regression, we assigned score to subjects for membership in Subtype 4 and 5, and then used the resultant continuous trait in the subsequent GWAS. In addition to the two subtypes, we also conducted a GWAS for the positive DSM5 diagnostic criterion count. Modulatory effects from four childhood environmental variables were considered in the association study. They were poor parental monitoring, unstable neighborhood, household drinking and illicit drug use, and household tobacco use. Subjects in the GWAS were genotyped using the Illumina OmniQuad microarray and imputed by referencing to the 1000 Genomes reference panel. This dataset consisted of 2,070 African Americans (AAs) and 1,570 European Americans (EAs), for which separate analysis was performed. The findings from GWAS were evaluated using a separate dataset (918 AAs and 1,382 EAs), the genotypes in which were obtained from both exome sequencing and imputation using also 1000 Genomes reference panel.

We observed, in GWAS, associations at genome-wide significant level from 13 independent genomic regions. For six of the identified genetic variants, data were available in the evaluation sample. The association for a variant *5:173935380* (at base pair position 173,935,380 in chromosome 5, p=1.23×10e-8), a SNP located in gene *LINC01411*, was successfully replicated (p=3.63×10e-3). Three additional variants, *2:202256694* (*TRAK2*), *1:15511771* (*TMEM51*) and *1:82524289* (*LPHN2*) remained associated at genome-wide significant level after meta-analysis. All above four associations were observed in AAs only.

.