Submitter Name: Roseann E. Peterson Submitted Email: roseann.peterson@downstate.edu

Expanded Empirical Population Assignments in Biobanks Empower Ancestrally Inclusive Genetic Studies of Substance Use Disorders

Madhurbain Singh[1,2], Amanda Elswick Gentry[1,3], Chris Chatzinakos[4,5,1,3], Bradley Todd Webb[6,1,3], Roseann E. Peterson[4,5,1,3]

[1]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University;
[2]Department of Human and Molecular Genetics, Virginia Commonwealth University;
[3]Department of Psychiatry, Virginia Commonwealth University;
[4]Department of Psychiatry and Behavioral Sciences, State University of New York - Downstate Health Sciences University;
[5]Institute for Genomics in Health, State University of New York - Downstate Health Sciences University;
[6]GenOmics, Bioinformatics, and Translational Research Center, RTI International.

The UK Biobank (UKB) is one of the largest data resources for genome-wide association studies (GWAS) of substance use disorders (SUDs). Most GWAS in the UKB are restricted to individuals of self-identified “White British” background to minimize heterogeneity and population stratification, excluding tens of thousands of individuals. However, advances in statistical methods allow the inclusion of relatively small cohorts of non-European ancestry in trans-ancestry GWAS to empower the discovery of genomic risk loci while enabling progress toward more equitable translational benefits. Here, we highlight the potential of trans-ancestry GWAS of SUDs using our ancestrally inclusive analytic pipeline. Applying genetic-based principal component analysis, we empirically assigned UKB samples (Application ID 30782) to six major population groups based on each individual’s multivariate distance from an expanded reference panel. As an example phenotype, we identified participants with an ICD diagnosis of nicotine dependence, yielding a total of 42381 cases and 253240 controls, representing a 12% gain in effective sample size over the “White British” subset, and comprising 39915+243414 European, 814+2752 Central/South-Asian, 637+2598 African, 314+1267 Admixed-Americas, 274+947 Middle-Eastern, and 109+771 East-Asian ancestry cases+controls, respectively. While the non-European groups have small sample sizes for ancestry-specific GWAS, these samples can be used for more inclusive, trans-ancestry GWAS – whether as joint mixed-effects GWAS or ancestry-stratified meta-analysis – and for contributing summary statistics to large-scale meta-analyses. Through this approach, we underscore the scientific and ethical rationale of trans-ancestry GWAS in biobanks, aiming to enhance representativeness and generalizability across the ancestry spectrum by maximizing inclusivity and addressing population stratification.