Name: Michael Garvin Email: garvinmr@ornl.gov

Long-Read Sequencing Identifies CYP2D6 Haplotype Associated With Frequent Cocaine Use

Michael R. Garvin[1], Daniel M. Rosenblum[2], Andrew W. Bergen[2], Stanley H. Weiss[2]

[1]Oak Ridge National Laboratory, [2]Rutgers New Jersey Medical School

We recently discovered a 3-SNP haplotype significantly associated with frequent cocaine use. The Genotype Expression Database shows this haplotype regulates expression of CYP2D6 in brain but not liver. The CYP2D6 locus is complex and known to harbor copy number variation (CNV), which is difficult to parse with standard sequencing. We used three novel approaches to characterize copy number and functional variants at CYP2D6. First, we analyzed raw intensity values from the genotyping array that indicated the 3-SNP haplotype is likely marking CNV. Next, we used long-read assemblies from the Human Pan Genome Project to reveal that the 3-SNP risk haplotype is linked to genomic features that could explain the association, including an interferon gamma regulatory site, amino acid changes in the CYP2D6 protein, and a splice site alteration adjacent to exon 4. Then, to verify these alterations in our subjects, we performed targeted long-read sequencing of the CYP2D6 locus in 19 subjects with 0, 1, or 2 copies of the risk allele and with varying frequencies of cocaine use. We confirmed our CNV prediction by demonstrating that individuals with no risk allele carried a single copy of CYP2D6, identical to the protein sequence in the human reference genome. All but one subject with two risk alleles carried a CYP2D6 haplotype comprising four amino acid changes and a splice site change. Our structural prediction using AlphaFold shows the splice site change may produce a novel CYP2D6 protein, which we posit is associated with frequent cocaine use in those of European ancestry.