Online supporting information for:

Gene x smoking interactions on human brain gene expression: Finding common mechanisms in adolescents and adults

Samuel Wolock,[1*] Andrew Yates,[2*] Stephen A. Petrill,[3] Jason W. Bohland,[4] Clancy Blair,[5] Ning Li,[1] Raghu Machiraju,[2,6] Kun Huang,[2,6,7] Christopher W. Bartlett[1,8]

[1]Battelle Center for Mathematical Medicine, Nationwide Children's Hospital, Columbus, OH; [2]Department of Biomedical Informatics, The Ohio State University College of Medicine, Columbus, OH; [3]Department of Psychology, The Ohio State University, Columbus, OH; [4]Department of Health Sciences, Boston University, Boston, MA; [5]Department of Applied Psychology, New York University, New York, NY; [6]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH; [7]The CCC Biomedical Informatics Shared Resource, The Ohio State University; [8]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, USA

*These authors contributed equally to this work

Running head: Brain eQTL G×E

Text mining for gene selection.

A text mining survey of the psychiatric genetics literature using FACTA+ (Tsuruoka et al., 2008) shows that up to 29% of the psychiatric literature includes discussion of the role of serotonin – the majority of those studies include the 5-HTTLPR polymorphism upstream of the serotonin transporter gene (*SLC6A4*) as the primary data for that discussion. Dopamine as a topic covers 21% of the literature with the functional variants in the following genes, *DAT1*, *COMT*, *DRD2*, *DRD4*, *DBH*, and *MAO-A* (by order of frequency), accounting for virtually all of the dopamine literature. The next largest block of published genetic variants, 13% of the literature, comprise functional variants within *BDNF*, *CYP450* and *APOE*. In total, 63% of the psychiatric genetics literature involves only 10 genes, all with known functional variants (either protein coding variants or eQTLs). Genes for this study were comprised of the top 100 genes most commonly discussed in the literature, and further included the remaining neurotransmitter receptor and metabolism genes in each family (acetylcholine, dopamine, GABA, glutamate and serotonin) for a total of 158 genes.

COLANTUONI gene expression

We downloaded the mRNA microarray expression data as the normalized, filtered, and $\log_2$ transformed expression matrix as prepared by (Colantuoni et al., 2011) from the online NCBI GEO database from GEO accession GSE30273, supplementary file GSE30272_RGna.n269.o30176.log2ratio.loess.MADout.KNNimp.txt.gz. We downloaded the sample phenotype annotations from the metadata describing each of the 269 GSM sample accessions organized under GSE30273 and the mRNA microarray probe annotations for the platform used, Illumina Human 49K Oligo array (HEEBO-7 set), as the GEO platform accession GPL4611.

Quality Control and Sample Selection. From the 269 samples in the NCBI dataset, we removed 109 samples with an age of less than 18 years or with "Unknown" smoking history values, leaving a total of *N*=160 samples. The dataset was previously normalized and confirmed by inspecting the distributions of the expression levels. Batches were defined by the concatenation of the variables "medical examiner office" and "brain bank source." Two batch categories (*N*=7 subjects) were too small for further processing and were dropped, as were nine individuals with race "Hispanic" or "Asian" (five and four samples, respectively), leaving the final *N*=144 samples (38 smokers and 106 non-smokers). Covariate correction was conducted with ComBat (Johnson et al., 2007), which corrects for categorical variables batch (see above), sex and race, while the correctBatchEffect routine from the *limma* library (Smyth, 2005) in the R statistical language was used for linear regression covariate correction of age, PMI, RIN and brain pH.

LIU gene expression

We downloaded the LIU (Liu et al., 2010) mRNA microarray expression data from the Stanley Genomics website (https://www.stanleygenomics.org/contact.html, free registration required). The data were generated using the Affymetrix Human Genome U133A Array (HG-U133A), and the unprocessed (raw) CEL files were stored in two compressed directories (study1_raw.zip and study2_raw.zip). We downloaded the updated mRNA microarray probe alignments and annotations for the Affymetrix HG-U133A platform from the University of Michigan Brainarray Custom CDF online database (Version 16.0.0, ENTREZG) (Dai et al., 2005) including hgu133ahsentrezgprobe_16.0.0.tar.gz and hgu133ahsentrezg.db_16.0.0.tar.gz .

Quality Control and Sample Selection. Using the normalization method SCAN-UPC (Piccolo et al., 2012) with the Brainarray probe annotations, we normalized 22,283 probes from 142 individuals, removing probes annotated as low quality or technical probes, leaving 12,079 probes. For whole array level quality control, we inspected both a boxplot and a histogram of sample array means of the SCAN-UPC "Universal Probability of expression Codes" (UPC) and removed five subjects with abnormally low UPC distributions (mean array UPC less than 0.22). Probe level data were filtered to remove features with median UPC scores indicating less than 20% confidence the transcript was expressed above background. A total of 53 samples missing genotypes, smoking status, or psychiatric diagnosis were dropped from further consideration. All subjects were greater than 18 years old and 84 had data on smoking status. Covariate correction was conducted as described for the COLANTUONI dataset, with the addition of the categorical variable "diagnosis." After quality control filtering there were 84 subjects (28 smokers and 56 non-smokers).

Genotype Data

We downloaded genotype data for the COLANTUONI dataset from dbGAP consisting of BeadExpress genotype calls from either the Illumina Human1M-Duov3 or the HumanHap650Yv3 arrays hybridized with cerebellar DNA. Affymetrix SNP 5.0 genotype calls for LIU were downloaded from www.stanleygenomics.org. SNP microarray genotyping for the LIU dataset was conducted on DNA extracted from the cerebellum of the same subjects with gene expression data, and calls were made using the BRLMM algorithm (Liu et al., 2010). These data were cleaned using standard procedures (Simmons et al., 2010). Briefly, using PLINK v1.07 (Purcell et al., 2007), SNPs in each of these files were

extracted for downstream analyses based on genotyping rate >= 95% for both subjects and SNPs, minor allele frequency (MAF) > 1%, and Hardy-Weinberg equilibrium $p<0.001$. The IBS and MDS routine in PLINK (Purcell et al., 2007) was used to create principal components from the genotype data, the top three of which were used as covariates to reduce the effects of SNP allele frequency differences purely due to population differences (Hou et al., 2011).

Imputation of SNPs on a common reference panel was conducted with MaCH for pre-phasing and minimac for the actual imputation step (Howie et al., 2012, Scott et al., 2007), with the exception that males do not require phasing on the X chromosome. Reference haplotypes were a combination of 1000 Genomes Project panels EUR and AFR as recommended (Chanda et al., 2012) for African American samples and EUR for the European ancestry samples. MaCH format 1000 Genomes Phase I v2 files were downloaded from http://www.sph.umich.edu/csg/yli/mach/download/. Imputed SNPs were filtered for MAF > 1% and $R^2 > 0.3$ to remove poorly imputable SNPs. Only SNPs within 1 Mb of the transcription start site (TSS) and transcription end site (TES) for the genes of interest were used in the eQTL analyses. These were further filtered to include only markers with at least three heterozygous subjects each in the smoking and non-smoking groups, Hardy-Weinberg equilibrium $p<0.001$, and genotyping rate >=95%. Once the SNPs from each dataset had been filtered individually, only those SNPs that were shared by both datasets were kept, yielding 405,875 SNPs.

Statistical Analysis

SNP×Smoking interactions. Linear regression was conducted with the R library Matrix eQTL (Shabalin, 2012) using the modelLINEAR_CROSS model. Gene expression

values were the dependent variable, while SNP and smoking were the independent variables. Covariates included age, sex, PMI and three principal components for ancestry (see above). The top 10,000 SNPs from the SNP×Smoking interaction analysis in each dataset were included in the meta-analysis. Within each dataset, a permutation *p*-value was obtained to correct for multiple testing of SNPs and genes in that dataset. The corrected *p*-values were used for meta-analysis. *P*-values were combined using the z-transform method weighted by sample size (Whitlock, 2005) as implemented in the survcomp R package (Haibe-Kains et al., 2008). Briefly, each *p*-value is converted to a standard normal deviate through the inverse normal distribution. The sum of the deviates, weighted by sample size, was divided by square root of the weights squared. This procedure is appropriate when all studies assume the same null hypothesis, as is the case here, with a uniform statistical procedure. For clarity of interpretation, additional correction of the *p*-values to assess genome-wide significance was accomplished by rescaling the meta-analysis output to be on the nominal scale post-correction. Using a strict Bonnferoni correction for multiple testing assuming 29 genome scans would require a threshold of $P< 1.7x10^{-9}$. However, this number is overly conservative since we only scanned 2.1% of the genome. Factoring this in, the appropriate threshold to correct for multiple testing is $8.1*10^{-8}$. Therefore, SNP-gene pairs were considered significant if their G×E meta-analysis *p*-value was less than 8.1×10^{-8} and if the direction of the effect was the same in both the COLANTUONI and LIU datasets. SNPs were considered to have different direction effects if the permutation *p*-value was less than 0.2 in both datasets and if the sign on the *t*-statistic was different between the two datasets.

Main effect of SNPs. Analysis of the effect of SNP on gene expression was conducted using the same analysis and permutation procedures but substituting modelLINEAR for the analysis model in Matrix eQTL (Shabalin, 2012). All subjects (*N*=186 for COLANTUONI, *N*=127 for LIU) were used in this analysis, including those for whom smoking status was unknown, and therefore the sample size for this analysis was different than for the analyses that included smoking status.

Main effect of smoking on gene expression. To generate the top differential expression results, we calculated *p*-values from a parametric F-test comparing a model corresponding to an effect of smoking status (smoking vs. not smoking) to the null model for each expression probe using the f.pvalue function in the *sva* Bioconductor v2.11 package (Gentleman et al., 2004). For each probe, we also compute false discovery rate (FDR) from the F-test *p*-values using the Benjamini–Hochberg technique (Benjamini and Hochberg, 1995), and we computed alternate *p*-values using Welch Two Sample *t*-tests.

Bayesian genetic modeling. The linear regression assumes an additive model for SNP and smoking effects. To assess if a non-additive model was more appropriate, genetic model parameters were estimated by the program Kelvin, a flexible platform for modeling complex genetics including quantitative traits including non-normal and truncated (censored) data, Gene×Gene, and G×E interactions (Bartlett and Vieland, 2005, Bartlett et al., 2007, Hou et al., 2012, Huang et al., 2007, Huang and Vieland, 2010). The quantitative trait likelihood includes a genotypic mean and variance for each of the three SNP genotypes, the allele frequency for the unobserved true underlying trait SNP genotype, and a linkage disequilibrium parameter to model the correlation between an observed (genotyped or imputed) SNP and the trait SNP. All parameters are estimated through

maximum likelihood by accepting the maximum value encountered during numerical integration of the trait parameters out of the likelihood when calculating the posterior probability that a SNP is associated with the trait (for more computational and statistical details see (Vieland et al., 2011).

SNP annotations

We used Build hg19/GRCh37 of the human genome as the reference sequence for both gene and SNP locations. RefSeq data was used for gene locations. SNP properties were annotated with SNPNexus (http://snp-nexus.org/) (Chelala et al., 2009, Dayem Ullah et al., 2012, Dayem Ullah et al., 2013), which provides information about the location of SNPs relative to nearby regulatory elements such as transcription factor binding sites, and other properties of the SNP such as predicted pathogenicity of amino acid changes. We processed "broad peaks" from ENOCDE ChIP-seq experiments performed on all cell and tissues classified as neutrally derived were downloaded and processed as bed files to assess the intersection with the location of significant eQTL SNPs.