Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Nikhil Ramgiri April 9, 2020

Integrative Prioritization of Genetic Loci for Nicotine Consumption

by

Nikhil Ramgiri

Rohan HC Palmer
Adviser

Emory University Department of Psychology

Rohan HC Palmer

Adviser

Daniel Weissman

Committee Member

Jingjing Yang

Committee Member

2020

Integrative Prioritization of Genetic Loci for Nicotine Consumption

By

Nikhil Ramgiri

Rohan HC Palmer

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Emory University Department of Psychology

2020

Abstract

Integrative Prioritization of Genetic Loci for Nicotine Consumption
By Nikhil Ramgiri

The consumption of nicotine products constitutes a serious public health concern, due to the substance's addictive properties and its potential to disrupt psychosocial functioning. Evidence from twin and molecular genetics studies strongly suggest that nicotine consumption, as a trait, adhere to a polygenic model. To date, traditional approaches to characterizing the genetics of polygenic traits, such as genome-wide association studies of nicotine/tobacco use disorders, have been limited in their ability to resolve genomic loci that contribute to the liability to misuse. Drug exposure paradigms in animal models provide us with a potentially useful source of cross-species gene expression data; the current study thus attempted to utilize transcriptomic data from nicotine/tobacco exposure studies in model organisms to better capture genetic variance in nicotine consumption in human populations. The following thesis addresses two primary objectives. Firstly, we construct and assess the viability of an integrative framework that leverages functional cross-species data to characterize the genetic underpinnings of nicotine consumption. Secondly, we determine whether regions of the genome localized by cross-species expression data can inform prediction of nicotine-related phenotypes in an independent human target sample. Our findings indicate significant enrichment of co-transcriptionally regulated loci identified via cross-species data; additionally, these loci carried significant predictive utility when applied to an independent human sample. Our research thus puts forth a promising approach towards unraveling polygenic variants involved in the neuro-molecular physiology of nicotine consumption and other drug use phenotypes.

Integrative Prioritization of Genetic Loci for Nicotine Consumption

By

Nikhil Ramgiri

Rohan HC Palmer

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Emory University Department of Psychology

2020

Acknowledgements

I would like to thank Dr. Rohan Palmer, Dr. Chelsie Benca-Bachman, and Dr. Spencer Huggett for their continued support and advice throughout the process of completing this honors thesis. Additionally, I would like to thank the members of my thesis committee – Dr. Jingjing Yang and Dr. Daniel Weissman – for their valuable insights and feedback. This honors thesis has contributed elements integrated into a manuscript that is currently pre-published to Biorxiv – "Cross-Species Integration of Transcriptomic Effects of Tobacco and Nicotine Exposure Helps to Prioritize Genetic Effects on Human Tobacco Consumption". As such, I would also like to acknowledge the corresponding authors on that manuscript for their creative efforts.

Table of Contents

Introduction 1

Background of Nicotine/Tobacco Dependence 1

Characterizing the Etiology of Nicotine/Tobacco Dependence 2

Study Aims 5

Methods 6

Assembling a Prioritized Subset 6

Genome Partitioning and Heritability Estimation 7

Mixed Linear Model Association Analysis 8

Polygenic Risk Score Modeling 9

Results 11

Estimation of SNP Heritability of Human Nicotine Consumption across Regions-of-Interest 11

Proportion of Gene SNPs Observed in UK Biobank MLMA P-value Distribution 13

Application of Genomewide Polygenic Effects and Gene Sets in Add Health Replication Sample 14

Discussion 16

References 21

List of Table and Figures References 27

Collection of Tables 28

Collection of Figures 34

Introduction

Background of Nicotine/Tobacco Dependence

Tobacco smoke, whether via first-hand or environmental inhalation, has long been implicated in deleterious health outcomes, chief among them being cancers of the lung, oral cavity/throat, and the liver[1]. Research delving into the chemical composition of tobacco smoke and its biological effects at the cellular level has identified components, including polycyclic aromatic hydrocarbons (PAHs), aromatic amines, and N-nitrosamines, among others, to be direct causal factors in tumorigenesis[2]. Metabolic activation of these compounds promotes the formation of DNA adducts, which covalently bind to DNA in pleural cells, interfering with replication and inducing mutant cell production[3]. Since nicotine is not among these identified carcinogens in tobacco products, contemporary marketing strategies by tobacco companies have shifted towards selling non-tobacco nicotine products, such as e-cigarettes, juul, and other vapor-producing products, in an effort to destigmatize nicotine as a harmful substance. Nicotine, however, can cause changes in neural organization, particularly in the brain's reward systems, and in psychomotor and cognitive processes via its ability to interact with naturally occurring nicotinic acetylcholine receptors (nAChRs)[4; 5]. By altering neural circuits, especially those comprising the dopaminergic systems of the midbrain, nicotine can elicit high potential for addiction, regardless of the form in which it is consumed[6]. The addictive properties of nicotine therefore increase the risk for individuals to engage in patterns of overconsumption and eventually experience a loss of psychosocial function.

Nicotine dependence has been observed to run in families beyond what can be explained by environmental factors. Heritability estimates for nicotine dependence obtained from twin study designs range from 40-70%, suggesting a high contribution of additive genetic elements to the

trait's total variance in the population[7]. In general, genetic effects on tobacco involvement have been shown to vary based on the phenotypic construct being studied. For instance, nicotine withdrawal heritability has been estimated at 54%[8], smoking initiation at 44%[9], and nicotine consumption as 80%[10]. These differences in heritability estimates across constructs suggest not only that nicotine involvement is a highly heritable phenotype, but also that many genetic elements exhibit widespread pleiotropic effects[11]. Considered in light of the extensive variation in expression of the trait within the population, it is likely that the genetic architecture of nicotine use and dependence adheres to an polygenic model, in which thousands of loci across the genome contribute to the trait of interest through interconnected gene regulatory networks[12]. Given this knowledge, the task of unearthing the genetic underpinnings of nicotine dependence has gained prime importance, particularly as sequencing technologies and large population registries have increased our access to available genotypic and phenotypic data. Identification of the specific genetic risk factors and causal pathways that collectively contribute to the manifestation of nicotine dependence in the population would allow for the implementation of prevention strategies tailored to individuals who carry increased risk within their genomes.

Characterizing the Etiology of Nicotine/Tobacco Dependence

Individual genetic markers, such as single-nucleotide polymorphisms (SNPs), have been the primary subject of modern approaches to characterizing genetic contributors to complex disorders, which are affected by many variants with small effect sizes. The current approach to capturing genetic risk underlying such complex disorders has been the use of genome-wide association studies (GWAS), which harness the statistical power afforded by very large sample sizes to identify reproducible SNPs that are associated with the phenotype of interest. While this

approach has done well in laying the groundwork for characterizing the genetics of complex traits, it is not without significant limitations. Firstly, GWAS designs carry the risk of being underpowered to detect causal variants, due to the strict significance threshold necessarily set for multiple comparisons. For example, Saccone and colleagues[13] indicated in 2007 that the SNP representing the CHRNB3 gene, which encodes for the β3 nicotine acetylcholine receptor (nAChR) subunit, produced a low enough statistical p-value to reach genome-wide significance. However, Gelernter and colleagues[14] found in 2015 that the gene cluster on chromosome 15 housing nicotinic receptor genes, including the same CHRNB3 gene, failed to reach genome-wide significance in their sample. This failure of GWAS designs to replicate significant results across studies, particularly for a gene encoding a subunit of an nAChR, long understood to mediate nicotine's addictive properties in the brain, underscores the need for extremely large sample sizes, often in the 10's to 100's of thousands, to avoid potential causal variants falling through the cracks. Recent efforts that have focused on meta-analyzing GWAS summary statistics, such as the large-scale association analyses of tobacco and alcohol use using the data from 1.2 million individuals aggregated across several genetic consortiums conducted by Liu and colleagues[15], have begun to combat this issue. Under normal circumstances however, such large sample sizes place severe constraints on available funding and become impractical without large-scale collaboration efforts or extensive computational resources. Secondly, significantly associated SNPs identified via GWAS study designs often carry small effect sizes, particularly in traits that adhere to a polygenic model, and are thus not always readily linked to discernable targets for pharmaceutical therapy. Because of these limitations, the translatable target loci that tend to emerge from drug dependence GWAS findings have been mostly restricted to receptors and metabolizing enzymes, leaving a large proportion of the variance unaccounted for.

One potential approach to increasing our power to detect causal loci contributing to liability for nicotine dependence is by using prioritized subsets of variants, based on prior experimental evidence. In 2008, Li et al. proposed an approach they termed "prioritized subset analysis" (PSA), in which they demonstrated that integrating prior information regarding relevant trait-associated loci into GWAS designs improved power to detect risk markers in a variety of model phenotypes[16]. Specifically, they found that the degree of power improvement increased the more number of identified risk markers were contained in the prioritized subset[16]. Additionally, recent work investigating the genetic architecture of various drug dependence syndromes has indicated that common variants contributing most to the expression of opioid, alcohol, and nicotine addiction phenotypes are not homogeneously spread across the genome, instead suggesting they could be enriched in certain regions[17-19]. Assessing these findings in the context of the omnigenic model of nicotine dependence, these regions would likely constitute loci involved in the pharmacokinetic and pharmacodynamic effects of nicotine. Since GWAS designs seem to capture pharmacokinetics (in the form of receptors and drug metabolizing enzymes) much better than they do pharmacodynamics, we could perhaps utilize the PSA to integrate prior evidence from both domains with traditional GWAS to better characterize genetic liability in the population.

To implement a prioritized subset approach in characterizing regions of potential enrichment, we need an appropriate source of evidence from which to construct the prioritized subset. The most relevant source of prior evidence that could be harnessed for genetically mediated effects of tobacco/nicotine on nicotine dependence would be differential gene expression data taken from brain tissue from a living, intact human. Exposing sample brain cells to acute or chronic administrations of nicotine would shed light on specific epigenetic and

expression changes that occur at the molecular level and would directly identify expression quantitative trait loci (eQTLs) most sensitive to the pharmacodynamic effects of the drug. Since such rich experimental data from living, intact human samples are scarce, another possible source of prior knowledge to be used in our prioritized subset lies in animal models of drug dependence.

Much like brain tissue from living, intact humans, we can use model organisms (*M. musculus, R. norvegicus, D. rerio,* etc.) to measure differential gene expression as a function of nicotine/tobacco exposure. Animal models provide a set of distinct advantages that make them ideal for use in this case. Firstly, they afford a degree of tight experimental control, in that environmental factors in the laboratory setting and genetic background (through the use of known inbred strains) can be closely monitored and manipulated by the experimenter. Secondly, exposure paradigms, in which pharmacokinetic, pharmacodynamic, and behavioral effects of given drug doses are measured, can be modeled in a manner that would be unethical to do in humans. Following identification of differentially expressed loci, phylogenetic similarities between animal and human genomes can be harnessed to produce a list of loci that can serve as the basis of a prioritized subset in characterizing the genetic architecture of nicotine dependence phenotypes in human populations.

Study Aims

The goal of the present study is two-fold. Firstly, we construct a novel integrative pipeline to determine whether functional gene expression data from model organism studies can be leveraged to streamline the process of characterizing the genetic underpinnings of the nicotine consumption phenotype. The viability of using animal models would be apparent if the

transcriptionally informed loci yielded by these studies can account for a significant proportion of the total genetic variance observed for the trait in a human sample (see Figure 1). We specifically examine nicotine consumption as a phenotype in order to reduce issues of translatability between the exposure paradigms used in the selected model organism studies and the corresponding trait in humans. Secondly, we seek to determine whether the regions of enrichment localized using our informed subset from animal models can inform prediction of nicotine consumption in an independent human target sample. We tested this hypothesis directly by constructing a polygenic risk score (PRS) using the effect sizes of markers housed in the enriched regions and assessing whether it captures a significant amount of the total variance of the phenotype in the independent sample.

Methods

Assembling a prioritized subset

To construct a prioritized subset of loci, we used the GeneWeaver ontological system, a repository of functional gene expression data from various model organisms with an accompanying suite of tools to allow for further gene-set analysis[20]. We conducted a query of the database to identify gene sets that originated from empirical studies in which a nicotine/tobacco exposure or consumption paradigm was implemented. In these experiments, mRNA was isolated from the animal's brain tissue following exposure/consumption and gene expression profiles were built from DNA microarray, RNA sequencing, or weighted gene co-expression network analysis (WGCNA) data (see Table 1 for full information on study paradigms).

The full list of genes from the GeneWeaver query was merged with the human genome (hg19/build 37) to produce an orthologous subset for use in further analyses. The empirical query

of the GeneWeaver database yielded a total of 923 genes differentially expressed as a function of nicotine exposure/consumption, 742 of which were matched onto the human genome. Of these, only 201 genes were found across more than one study, suggesting the pharmacodynamic effects of nicotine are highly heterogeneous (see Figure 2 for distribution of replication across studies).

Genome Partitioning and Heritability Estimation

To determine the relative effect on the phenotype by the genes compiled from GeneWeaver, we partitioned the length of the genome into three regions of interest. The "gene" region encompassed all genes shown to differentially expressed as a function of nicotine exposure in the empirical animal literature. The flanking "buffer" regions encompassed the base pairs directly upstream of the 5' end and directly downstream of the 3' end of each of the genes. Five buffer lengths (5kB, 10kB, 25kB, 35kB, and 50kB) were considered in an effort to capture any potential variation imparted by transcription factor binding sites (TFBS) and other regulatory elements, whose exact positions are variable and unknown. The "all other variants" regions encompassed all variants that did not fall into either the "gene" of the "buffer" regions (see Figure 3 for diagrammatic depiction of genome partitioning).

The human analytical sample used for heritability estimation was drawn from individuals in the UK Biobank[21] who reported being either a current or former smoker (n = 123,844). The nicotine consumption phenotype was described as the number of cigarettes smoked per month. The total sample was split into three constitutionally equivalent folds (n_1 = 41,263, n_2 = 41,368, n_3 = 41,213) to allow for computational efficiency and to demonstrate robustness of findings via replication of results across folds.

The genetic data of the individuals in each fold were fit to a multivariate model via GCTA-GREML[22] to evaluate the relative contribution of each region-of-interest to the expression of the phenotype. Sex, testing site location, age, and age^2 were controlled for in this analysis. Enrichment values (E) were calculated for the gene regions, each of the varying buffer lengths, and the "all others" regions to determine whether the observed component-heritability estimates were greater than what would be expected by chance, given the total genetic variance for the nicotine consumption phenotype in the sample and the 4.6 million SNP markers used in the analysis. Calculated in this manner, the enrichment values represent the ratio of the observed effect size to the effect size expected by agnostic, unbiased selection of loci across the genome.

$$\text{Expected h2SNP} = \frac{\#\text{SNPs}ROI \text{ x Obs}h^2_{\text{SNP\_Total}}}{\#\text{SNPs}Total}$$

Mixed Linear Model Association Analysis

To highlight the potential utility of harnessing expression data from the animal model literature alongside traditional agnostic GWAS designs, we carried out a mixed linear model association (MLMA) analysis, implemented in GCTA via the MLMA-LOCO[23] option. The MLMA-LOCO assessed the association between genotyped SNP markers (n_{SNPs} = 4656938) in the UK Biobank sample of individuals (n = 123,844) and the consumption phenotype (measured as cigarettes smoked per month).

SNP markers were then divided into bins based on p-values yielded by the cigarettes-per-month MLMA-LOCO analysis. Bins ranged from the genome-wide significance threshold of p <= 5 x 10^{-8} to p = 1. A distribution was then created to express the proportion of SNPs in each bin found in the prioritized gene regions identified from the animal expression data.

Polygenic Risk Score Modeling

In order to inform prediction of nicotine-related phenotypes in an independent sample, we constructed a series of polygenic risk score (PRS) models generated from the GWAS summary statistics (p-values and OLS regression beta-coefficients) produced by the MLMA-LOCO in the full UK Biobank sample. Using a reference sample from the 1000 Genomes Project[24], effect sizes of variants were adjusted based on linkage disequilibrium patterns using the SBLUP method[25]. By re-estimating SNP effects by converting them into best linear unbiased predictors, the SBLUP method has demonstrated improved prediction accuracy over other common methods of building polygenic risk scores[26].

The independent target sample was composed of European-Americans with genotypic data at Wave IV of the Add Health data project[27] (N = 4102, mean age = 28.9, SD = 1.7; proportion male = 0.469). The Add Health data project is a longitudinal study that has collected survey data on various health, economic, physiological, and psychosocial variables from a sample of individuals, beginning at grades 7-12 through adulthood, for a total of four waves[27]. The same phenotype we explored in the UK Biobank heritability estimation analysis – consumption (measured as cigarettes per day/CPD and referred to as the primary phenotype for the rest of this manuscript) – was examined in this sample, along with four other smoking-related phenotypes contained in the Add Health dataset. The total number of individuals in the full genotyped sample who answered the CPD item in the Add Health Wave IV survey was 1667 (mean age = 28.9, SD = 1.7; prop. male = 0.506). The "frequency of use" item, asked to those who claimed to be a current or former smoker, was described as the number of days per month that the individual claimed to have smoked at least one cigarette (n = 1673; mean age = 28.9, SD = 1.7; prop. male = 0.507). The "age of first use" item, asked to all participants, was described as

the age at which the individual first smoked a whole cigarette (n = 2992; mean age = 28.9, SD = 1.7; prop. male = 0.485). The "age of initiation" item, asked to self-identified current or former smokers, was described as the age at which the individual first began smoking cigarettes on a consistent basis (n = 2158; mean age = 28.9, SD = 1.7; prop. male = 0.475). The "Fagerström index", compiled for self-identified current or former smokers, was a 10-point scale, meant to quantify nicotine dependence, that aggregated several items concerning cigarette craving, consumption, frequency, and heaviness of use (n = 2436; mean age = 28.9, SD = 1.7; prop. male = 0.475). For each phenotype, age and sex were regressed out as covariates, and the phenotype residuals were extracted for further analyses. Determination of how risk scores generated from summary statistics of nicotine consumption generalizes to other nicotine-related phenotypes would provide us with insights into the degree of overlap in genetic architecture among aspects of nicotine use.

The first model harnessed all the SNP variants (n_{SNPs} = 4656938) contained in the GWAS summary statistics produced by the UK Biobank consumption MLMA-LOCO (Model 1). Risk scores generated from SBLUP-adjusted variants from the base sample were regressed onto each phenotype in the target sample and assessed for strength of association (β_1 estimate) and variance explained (partial R_2). The first six principal components extracted from the genetic data of the Add Health sample were also included as covariates in all regression models to account for any confounding, which was also minimized using strict population homogeneity procedures described elsewhere[28]. A second model (Model 2), further partitioned the observed polygenic effect of all SNPs using the aforementioned regions of interest similarly examined in UKB (i.e., gene, "10kB buffer", and "all other variants"). These PRS were treated as distinct predictors when regressed collectively onto each phenotype in Add Health. The 10kB buffer was chosen

because this was the largest buffer size used in our UK Biobank models[29] before a drop in enrichment was observed. The variance in the phenotype accounted for by the risk scores of gene and buffer regions would indicate how well our findings from the UK Biobank analysis translate to a separate sample. As was done with the heritability estimations in the UK Biobank analyses, enrichment values were calculated for each region of interest to determine whether the observed effect of the corresponding PRS was greater than the value we would expect by chance. To calculate enrichment for a particular component, we randomly sampled, without replacement, the same number of SNPs housed in that region of interest from all markers across the genome 1000 times. We then used the selected SNPs to create corresponding risk scores and regressed each onto the phenotype to generate a distribution of standardized beta-values. Squaring these beta-values would subsequently provide a distribution of R^2-values; the mean of this sampling distribution would thus constitute the expected variance in the phenotype explained by that region of interest if markers were culled at random across the genome. Enrichment can then be assessed by the ratio of the observed to the expected effect sizes. A distribution p-value was also determined for each region of interest based on the proportion of permuted standardized beta values that were as much or more extreme than the observed β_1 estimate. In this sense, the distribution p-value represented the probability of obtaining the observed β_1 estimate given the distribution of permuted standardized beta values.

Results

Estimation of SNP Heritability of Human Nicotine Consumption Across Regions-of-Interest

The total R^2 accounted for by each model, in which the three regions of interest (gene, buffer, and "all others") were fit to the phenotypic data of the individuals, represented the

estimate of the total SNP heritability (i.e. the effect of additive genetic elements across the entire genome – h^2_{SNP}). The total h^2_{SNP} of nicotine consumption in the UK Biobank sample ranged from 7.5% to 9.5%, depending on the fold of individuals examined (see Table 2 for all estimates, as well as h^2_{meta} values). SNP variants housed in the prioritized genes (gene region-of-interest) accounted for approximately 0.2% to 0.4% of the total observed variance in nicotine consumption (Table 2). Variants around the prioritized genes (buffer region-of-interest) accounted for approximately 0.4% to 3.1% of the variance, while variants across the remainder of the genome (all others region-of-interest) accounted for 5.0% to 7.8% (Table 2). Interestingly, the R^2 of the buffer components across all models was greater than that of the gene components, despite the smaller buffer lengths housing fewer SNPs than the gene region.

The variance accounted for by the gene region-of-interest demonstrated little association with buffer length, as the relatively small changes in R^2 observed across models do not correlate with increasing buffer length (Table 2). However, a notable trend can be seen in the variance explained by the buffer and "all others" regions. As represented in Figure 4, the variance in nicotine consumption captured by the buffer region dramatically increases as buffer length in the models increases from 5kB to 50kB (0.4% to 3.1%, Table 2). Moreover, the variance captured by the "all others" region decreases with increasing buffer length (Figure 4), indicating that markers contributing to the phenotype are being repositioned into the buffer component of the model. This finding is consistent with previous work correlating variance explained with length of DNA, implicating a polygenic model in these data[30].

Significant enrichment was seen in variants clustered within genes (gene region-of-interest) and around genes (buffer region-of-interest) in nearly all models examined (Table 3). No enrichment was seen in the "all others" component of any of the six models (Table 3).

Enrichment values calculated for the gene region indicated that the observed effect contributed by our prioritized gene set was nearly twice as large as would be expected by randomly sampling the same number of SNPs across the length of the genome. The buffer region showed even greater values of enrichment (2.0 to 21.4) than the gene region, a result that was consistent with our earlier observation of the buffer region yielding a greater h^2_{SNP} estimate than the gene region despite the smaller buffer lengths housing fewer markers. While our earlier heritability estimation results demonstrated the effect size of the buffer component increased with increasing buffer length, we can also see an exponential decay in enrichment value at buffer lengths greater than 10kB (Figure 5). This finding would suggest that trait-associated variants are more enriched close to genes likely to be undergoing transcriptional regulation in both humans and mice.

Proportion of Gene SNPs Observed in UK Biobank MLMA P-value Distribution

The association analysis using the full sample of 123,844 current and former smokers identified in the UK Biobank dataset largely confirmed regions of the genome associated with cigarette consumption previously identified by the large-scale meta-analysis conducted by Liu and colleagues[15]; this was expected as UKB contributed a majority of samples to the paper. We identified 770 signals that reached genome-wide significance ($p < 5 \times 10^{-8}$), most of which were clustered on chromosomes 15, 19, 8, 7, 4, 3, and 1 (see Figure 6 for Manhattan plot).

Across the p-value distribution produced by the MLMA analysis, SNP markers from the prioritized genes were present in the highest proportion (0.074) among the variants found in the genome-wide significant bin ($p <= 5 \times 10^{-8}$) (Figure 7). The proportion of SNPs from the prioritized genes found in each bin gradually declined with increasing p-value, with the steepest

drop-offs occurring from $p <= 5 \times 10^{-6}$ to $p <= 5 \times 10^{-5}$ and from $p <= 5 \times 10^{-5}$ to $p <= 5 \times 10^{-4}$ (Figure 7).

Application of Genomewide Polygenic Effects and Gene Sets in Add Health Replication Sample

Model 1accounted for approximately 3.5% of the total variance in the primary phenotype ($R^2 = .035$, 90% CI [.026, .044], $p < .0005$) (Table 4). This suggested that consumption was under additive genetic influences in the Add Health study. The "all variants" PRS used in Model 1 was found to be significantly associated with CPD ($\beta_1 = .174$, 95% CI [.127, .221], $p < .0005$) and explained a large majority of the variance in nicotine consumption captured by the full model (partial $R^2 = .030$) (Tables 4 and 5).

Among the other four smoking-related phenotypes examined, the "all variants" PRS was found to be significantly associated with smoking frequency ($\beta_1 = .122$, 95% CI [.073, .171]), age of initiation ($\beta_1 = -.079$, 95% CI [-.131, -.028]), and the Fagerström Test of nicotine dependence ($\beta_1 = .164$, 95% CI [.115, .213]) (Table 5). The risk score did not inform prediction of age of first use.

Model 2, in which the total genomewide effect was partitioned into three separate polygenic scores, accounted for approximately 3.8% of the total variance in the primary phenotype ($R^2 = .038$, 90% CI [.029, .047], $p < .0005$) (Table 6). The polygenic scores generated from the gene region ($\beta_1 = .079$, 95% CI [.025, .132], $p = .004$) and the "all others" region ($\beta_1 = .164$, 95% CI [.116, .212], $p < .0005$) were significantly associated with the primary phenotype (Tables 6 and 7). Interestingly, the PRS generated from the 10kB buffer was not found to be associated with the primary phenotype in this sample.

The PRS generated from the gene region explained a small proportion of the total variance in the primary phenotype in the Add Health sample (R^2 = .0062) (Table 7), a magnitude that lies relatively in line with the h^2_{SNP} estimates for nicotine consumption gleaned from the gene region in the UK Biobank sample. Along with CPD, the PRS generated from gene region variants was significantly associated with smoking frequency (β_1 = .068, 95% CI [.013, .124], partial R^2 = .0046) and the Fagerström index (β_1 = .059, 95% CI [.004, .115], partial R^2 = .0035) (Table 7). The risk score generated from the "all others" region, was associated with CPD, smoking frequency, the Fagerström index, and age of initiation (β_1 = -.071, 95% CI [-.123, -.018], partial R^2 = .0050) (Table 7).

Enrichment calculation to determine whether the observed variance in the primary phenotype explained by the PRS constructed from our prioritized gene set was greater than expected by agnostic sampling of the same number of SNPs from across the genome illustrated significant enrichment in the gene region (E = 5.88, p = .02) (Table 7). An expected R^2 value was similarly calculated for the buffer region based on the number of SNPs housed in the 10kB buffer; however, since the observed R^2 found for the PRS generated from this region was not significantly different than zero, enrichment was not significant (Table 7). Despite a significant association between the "all others" PRS and the primary phenotype, no significant enrichment was found for this region (E = 0.90, p = 0.61) (Table 7).

The pattern of enrichment seen in smoking frequency mirrors that of the pattern seen in consumption, with the gene region exhibiting significant enrichment (E = 4.22, p = .031), the buffer region's enrichment fixed to zero, and the "all others" region demonstrating no significant enrichment (E = 0.90, p = .451) (Table 7). Regressions of the partitioned risk scores onto the Fagerström index phenotype also yielded a similar pattern, with the enrichment seen in the gene

region approaching statistical significance (E = 3.00, p = .059) and no significant enrichment in the "all others" region (E = 0.83, p = .305) (Table 7).

Discussion

Our findings indicate that cross-species expression data harnessed from model organism studies are a viable resource that can be used to improve our ability to characterize the genetic underpinnings of drug dependence phenotypes, such as nicotine consumption. Variants selected based on the *a priori* set of genes were determined to be significant contributors to smoking using only a third of the UK Biobank sample. The total h^2_{SNP} meta estimate of ~8% for nicotine consumption in the UK Biobank smoker sample corroborates estimates found in other studies examining similar phenotype[15; 18]. Partitioning this total genetic variance demonstrated that variants in and around genes shown to be differentially expressed as a function of nicotine exposure in model organisms account for a greater proportion of the total trait variance in the UK Biobank sample than would be expected by chance. The relatively large percentage of the genetic variance (4.2% to 39.5%, depending on buffer length considered) attributable to these regions, presumably involved in mRNA transcription, suggests that much of the heritability of nicotine consumption is driven by neuro-epigenetic changes in the brain upon exposure to the pharmacodynamic effects of the drug. Indeed, studies examining the neurobiology of other drugs have concluded that their biochemical and behavioral effects on the individual are induced by changes in gene expression levels in trait-relevant neurons in the central nervous system[31; 32]. Taken in light of the typical pathway of development towards drug dependence, it can be inferred that individuals with a higher genetic risk of excess nicotine use are more susceptible to a physiological response at the molecular level upon exposure that precipitates increased

susceptibility to greater future consumption[33]. Given that we were able to discern sizable neuro-molecular associations between our prioritized set of loci and the phenotype of interest, it stands to reason that leveraging functional data across species alongside traditional genome-wide association designs would enhance our ability to identify causal markers and neurochemical pathways related to nicotine consumption. The genetic loci supplied by cross-species exposure studies could serve to elucidate more of the specific pharmacodynamic mechanisms of nicotine in the brain; however, many of these loci tend to become buried under the stringent genome-wide significance threshold placed on human GWAS discovery findings. As evidenced by the proportions of SNPs from our prioritized subset found at various levels of the p-value distribution of our UK Biobank association analysis, our prioritized subset, built from animal expression data, draws from both significant and non-significant sources of variation to aggregate small effect sizes across the genome. Given the enrichment seen in our localized regions, it can be posited then that the most robust model attempting to characterize the genetic variance seen in nicotine consumption would be one that integrates cross-species expression data with genome-wide significant findings.

The second aim of the present study was to assess the translatability of the effects of the enriched regions in predicting nicotine use phenotypes in an independent sample of individuals. Using the SBLUP method of constructing polygenic risk scores, we found that a risk score created from the weighted sum of all SNPs genotyped in our Add Health target sample could account for ~3.5% of the variance in cigarettes smoked per day. This value is in line with the R^2 estimate determined by Liu et al[15], who created a polygenic risk score via the LDPred method using nearly identical discovery and target samples. We were also able to see that the risk score, which was generated from the summary statistics of a CPD GWAS, was able to capture

significant proportions of the variance in other nicotine use phenotypes, including smoking frequency, age of initiation, and the Fagerström index of nicotine dependence. This finding suggests that, while nicotine dependence remains a fundamentally multidimensional phenotype[18], the molecular pathways underlying each construct may still possess substantial genetic overlap.

Creating partitioned risk scores based on regions of enrichment localized by cross-species expression data demonstrated that our prioritized gene set carried significant predictive value, accounting for a larger proportion of the trait variance in the primary phenotype than would be expected by chance. The translatability of this gene set from the UK Biobank sample to the independent Add Health sample augments the viability of utilizing functional data across species in conjunction with genome-wide findings to better capture genetic variance in similar drug use phenotypes. In evaluating the application of the risk score generated from our prioritized gene set to other nicotine use phenotypes, the PRS managed to capture a similar proportion of the variance in smoking frequency as it did in consumption. However, it captured notably less variance in the Fagerstrom index, and it failed to capture any variance in age of initiation. These differences in partial R^2 indicate that our prioritized subset was specific to loci involved in the neuro-molecular changes associated with nicotine consumption and was therefore not wholly translatable to other nicotine-related constructs. Similar observations have been made with respect to alcohol consumption and problems stemming from use in the UK Biobank[34], as well as in the twin literature with respect to generalized drug dependence[35]. Given that the cross-species paradigms used to assemble our prioritized subset did not model the behavioral criteria evident in the Fagerstrom index and in age of initiation, the lack of direct applicability is perhaps not entirely surprising.

The findings of the present study should be interpreted in the context of some limitations. Firstly, the body of cross-species expression data generated through drug exposure remains rather narrow, limiting our understanding of the functional consequences of nicotine consumption to the results of available DNA microarray studies. As such, we cannot say for certain whether the pharmacodynamic effects of the nicotine exposure paradigm are the same across species. Additional work using post mortem human brains is necessary. Additionally, changes in gene expression levels as a result of exposure may differ across brain regions. While the integrative approach we have put forth has demonstrated promise and a necessary proof of concept, these considerations must be addressed by increasing the volume of model organism literature to maximize its utility. Secondly, we did not consider the specific regulatory elements involved in the differential expression of corresponding genes. We attempted to address this limitation by modeling flanking buffer regions of varying lengths, with the expectation that these regions housed transcription factor binding sites, upstream and downstream cis-acting enhancers and silencers, and methylation sites. While we were able to observe significant enrichment in these regions in the UK Biobank sample, the lack of association between risk scores generated from these regions and nicotine use phenotypes in an independent sample suggests that this method of modeling regulatory elements is not entirely generalizable and may require additional steps to identify and model the effects of regulatory elements by cell type (e.g., Hi C coupled MAGMA[36]). It remains to be seen to what degree regulatory mechanisms for specific loci are conserved across species.

In conclusion, our research puts forth a promising approach towards unraveling polygenic variants involved in the neuro-molecular physiology of drug use phenotypes, such as nicotine consumption. We show that incorporating *a priori* evidence from cross-species expression data

into traditional genome-wide findings in human populations can not only capture a greater share of the genetic variance in the trait, but also potentially offer improved clinical risk prediction. This integrative framework therefore represents a worthwhile approach to characterizing the genetic underpinnings of nicotine consumption.

References

1. Hecht, S.S. (1999). Tobacco Smoke Carcinogens and Lung Cancer. JNCI: Journal of the National Cancer Institute 91, 1194-1210.

2. Hoffmann, D.H.I. (1997). THE CHANGING CIGARETTE, 1950-1995. Journal of Toxicology and Environmental Health 50, 307-364.

3. Geacintov, N.E., Cosman, M., Hingerty, B.E., Amin, S., Broyde, S., and Patel, D.J. (1997). NMR Solution Structures of Stereoisomeric Covalent Polycyclic Aromatic Carcinogen−DNA Adducts: Principles, Patterns, and Diversity. Chemical Research in Toxicology 10, 111-146.

4. Changeux, J.-P., Edelstein, S., and Edelstein, S.J. (2005). Nicotinic acetylcholine receptors: from molecular biology to cognition.(Odile Jacob Publishing Corp).

5. Besson, M., Granon, S., Mameli-Engvall, M., Cloëz-Tayarani, I., Maubourguet, N., Cormier, A., Cazala, P., David, V., Changeux, J.-P., and Faure, P. (2007). Long-term effects of chronic nicotine exposure on brain nicotinic receptors. Proceedings of the National Academy of Sciences 104, 8155-8160.

6. GRENHOFF, J., ASTON-JONES, G., and SVENSSON, T.H. (1986). Nicotinic effects on the firing pattern of midbrain dopamine neurons. 128, 351-358.

7. Uhl, G.R., Liu, Q.-R., Drgon, T., Johnson, C., Walther, D., and Rose, J.E.J.B.G. (2007). Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. 8, 10.

8. Xian, H., Scherrer, J.F., Madden, P.A.F., Lyons, M.J., Tsuang, M., True, W.R., and Eisen, S.A. (2003). The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. Nicotine & Tobacco Research 5, 245-254.

9. Vink, J.M., Willemsen, G., and Boomsma, D.I.J.B.G. (2005). Heritability of Smoking Initiation and Nicotine Dependence. 35, 397-406.

10. Maes, H.H., Sullivan, P.F., Bulik, C.M., Neale, M.C., Prescott, C.A., Eaves, L.J., and Kendler, K.S. (2004). A twin study of genetic and environmental influences on tobacco initiation, regular tobacco use and nicotine dependence. Psychological Medicine 34, 1251-1261.

11. Visscher, P.M., and Yang, J. (2016). A plethora of pleiotropy across complex traits. Nature Genetics 48, 707-708.

12. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177-1186.

13. Saccone, S.F., Hinrichs, A.L., Saccone, N.L., Chase, G.A., Konvicka, K., Madden, P.A.F., Breslau, N., Johnson, E.O., Hatsukami, D., Pomerleau, O., et al. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. Human Molecular Genetics 16, 36-49.

14. Gelernter, J., Kranzler, H.R., Sherva, R., Almasy, L., Herman, A.I., Koesterer, R., Zhao, H., and Farrer, L.A. (2015). Genome-Wide Association Study of Nicotine Dependence in American Populations: Identification of Novel Risk Loci in Both African-Americans and European-Americans. Biological Psychiatry 77, 493-503.

15. Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature Genetics 51, 237-244.

16. Li, C., Li, M., Lange, E.M., and Watanabe, R.M. (2008). Prioritized subset analysis: improving power in genome-wide association studies. Hum Hered 65, 129-141.

17. Brick, L.A., Micalizzi, L., Knopik, V.S., and Palmer, R.H.C. (2019). Characterization of DSM-IV Opioid Dependence Among Individuals of European Ancestry. Journal of Studies on Alcohol and Drugs 80, 319-330.

18. Bidwell, L.C., Palmer, R.H.C., Brick, L., McGeary, J.E., and Knopik, V.S. (2016). Genome-wide single nucleotide polymorphism heritability of nicotine dependence as a multidimensional phenotype. Psychological medicine 46, 2059-2069.

19. Palmer, R.H.C., Brick, L.A., Chou, Y.-L., Agrawal, A., McGeary, J.E., Heath, A.C., Bierut, L., Keller, M.C., Johnson, E., Hartz, S.M., et al. (2019). The etiology of DSM-5 alcohol use disorder: Evidence of shared and non-shared additive genetic effects. Drug and Alcohol Dependence 201, 147-154.

20. Baker, E.J., Jay, J.J., Bubier, J.A., Langston, M.A., and Chesler, E.J. (2012). GeneWeaver: a web-based system for integrative functional genomics. Nucleic Acids Research 40, D1067-D1076.

21. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209.

22. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88, 76-82.

23. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M., and Price, A.L. (2014). Advantages and pitfalls in the application of mixed-model association methods. Nature genetics 46, 100-106.

24. Delaneau, O., Marchini, J., McVean, G.A., Donnelly, P., Lunter, G., Marchini, J.L., Myers, S., Gupta-Hinch, A., Iqbal, Z., Mathieson, I., et al. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nature Communications 5, 3934.

25. Maier, R.M., Zhu, Z., Lee, S.H., Trzaskowski, M., Ruderfer, D.M., Stahl, E.A., Ripke, S., Wray, N.R., Yang, J., Visscher, P.M., et al. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nature Communications 9, 989.

26. Robinson, M.R., Kleinman, A., Graff, M., Vinkhuyzen, A.A.E., Couper, D., Miller, M.B., Peyrot, W.J., Abdellaoui, A., Zietsch, B.P., Nolte, I.M., et al. (2017). Genetic evidence of assortative mating in humans. Nature Human Behaviour 1, 0016.

27. Harris, K.M., and Udry, J.R. (2018). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]. In. (Carolina Population Center, University of North Carolina-Chapel Hill [distributor], Inter-university Consortium for Political and Social Research [distributor].

28. Brick, L.A., Keller, M.C., Knopik, V.S., McGeary, J.E., and Palmer, R.H.C. (2019). Shared additive genetic variation for alcohol dependence among subjects of African and European ancestry. Addict Biol 24, 132-144.

29. Palmer, R.H.C., Benca-Bachman, C.E., Bubier, J.A., McGeary, J.E., Ramgiri, N., Srijeyanthan, J., Huggett, S., Yang, J., Visscher, P., Yang, J., et al. (2019). Cross-Species Integration of Transcriptomic Effects of Tobacco and Nicotine Exposure Helps to Prioritize Genetic Effects on Human Tobacco Consumption. bioRxiv, 2019.2012.2023.887083.

30. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics 43, 519-525.

31. Evangelou, E., Gao, H., Chu, C., Ntritsos, G., Blakeley, P., Butts, A.R., Pazoki, R., Suzuki, H., Koskeridis, F., Yiorkas, A.M., et al. (2019). New alcohol-related genes suggest shared genetic mechanisms with neuropsychiatric disorders. Nature Human Behaviour 3, 950-961.

32. Gelernter, J., Sun, N., Polimanti, R., Pietrzak, R.H., Levey, D.F., Lu, Q., Hu, Y., Li, B., Radhakrishnan, K., Aslan, M., et al. (2019). Genome-wide Association Study of Maximum Habitual Alcohol Intake in >140,000 U.S. European and African American Veterans Yields Novel Risk Loci. Biological Psychiatry 86, 365-376.

33. Sharp, B.M., and Chen, H. (2019). Neurogenetic determinants and mechanisms of addiction to nicotine and smoked tobacco. European Journal of Neuroscience 50, 2164-2179.

34. Johnson, E.C., Sanchez-Roige, S., Acion, L., Adams, M.J., Bucholz, K.K., Chan, G., Chao, M.J., Chorlian, D.B., Dick, D.M., Edenberg, H.J., et al. Polygenic contributions to alcohol use and alcohol use disorders across population-based and clinically ascertained samples. Psychological Medicine, 1-10.

35. Palmer, R.H.C., Button, T.M., Rhee, S.H., Corley, R.P., Young, S.E., Stallings, M.C., Hopfer, C.J., and Hewitt, J.K. (2012). Genetic etiology of the common liability to drug dependence: evidence of common and specific mechanisms for DSM-IV dependence symptoms. Drug and alcohol dependence 123 Suppl 1, S24-S32.

36. Sey, N.Y.A., Hu, B., Mah, W., Fauni, H., McAfee, J.C., Rajarajan, P., Brennand, K.J., Akbarian, S., and Won, H. (2020). A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. Nature Neuroscience.

List of Table and Figure References

Table 1: Identification of Empirical Studies Used to Assemble a Prioritized Subset for Further Analyses

Table 2: Estimated SNP-heritability for each Region of Interest in GCTA-GREML Multivariate Model

Table 3: Calculated Enrichment Values of Model Components

Table 4: Polygenic Model 1 Standardized Results for Cigarettes Per Day

Table 5: Strength of Associations and Variance Explained by PRS of All Variants

Table 6: Polygenic Model 2 Standardized Results for Cigarettes Per Day

Table 7: Regression Coefficients and Variance Explained by Partitioned Risk Scores

Figure 1: Theoretical integrative genomics approach to assessing translatability of expression data from model organisms to capturing genetic variance of nicotine consumption in human populations.

Figure 2: Distribution of genes identified via GeneWeaver query.

Figure 3: Visualization of each region-of-interest used as a model component in statistical analyses.

Figure 4: Visualization of partial R_2 accounted for by regions of interest across various multivariate models.

Figure 5: Enrichment decay of flanking buffer region seen with increasing buffer size.

Figure 6: Manhattan plot of UK Biobank MLMA for nicotine/tobacco consumption.

Figure 7: Proportion of SNPs across *successive bins of* UK Biobank GWAS p-value distribution that are found in the prioritized subset of genes obtained from animal model expression data.

Collection of Tables

Table 1. Identification of Empirical Studies Used to Assemble a Prioritized Subset for Further Analyses

Author(s)	GeneWeaver ID	Model Organism	Nicotine Consumption/Exposure Paradigm	Experimental Design	Brain Region	Number of Genes Contributed
Chen et al.	GS87128	Mus musculus	Subcutaneous acute nicotine treatment (expression changes measured at time-points of 1, 2, 4, and 6 hrs)	Microarray Analysis, WGCNA	VTA	184
Polesskaya et al.	GS14885	Rattus norvegicus	Subcutaneous chronic nicotine treatment (at ages p25, p35, p45, and p55)	Microarray Analysis, qRT-PCR, Principle Cluster Analysis	PFC, Ventral Striatum, Hippo.	66
Lee et al.	GS225897, GS225899, GS225900	Mus musculus	Intravenous nicotine self-administration	Microarray Analysis, qRT-PCR, WGCNA	Medial and Lateral Habenula	40
Wang et al.	GS14888, GS14889, GS14890, GS14891, GS14892, GS14893	Mus musculus	Nicotine administration in drinking water in two selectively bred mouse strains	Microarray Analysis, qRT-PCR, WGCNA	Amygdala, Hippo., nAcc, PFC, VTA	651
Kily et al.	GS14902, GS14903	Danio rerio	Nicotine-induced conditioned place preference	Microarray Analysis, qRT-PCR	Whole Brain	158
Sharp et al.	GS128167	Rattus norvegicus	Chronic nicotine self-administration	Microarray Analysis, RT-PCR	nAcc	188
Piechota et al.	GS355715	Mus musculus	Gene-expression changes (measured at time-points of 1, 2, 4, and 8 hrs) following acute nicotine injection	Microarray Analysis, qRT-PCR, WGCNA, In situ hybridization, Western blotting	Striatum	121

Note: GeneWeaver IDs can be used to review the full complement of genes supplied by each study

Table 2. Estimated SNP-heritability for each Region of Interest in GCTA-GREML Multivariate Model

Model component	*F1 h^2_{SNP}*	*F1 SE*	*F2 h^2_{SNP}*	*F2 SE*	*F3 h^2_{SNP}*	*F3 SE*	*h^2_{meta} (95% CI)*	*% total h^2_{meta}*
ROI - Genes								
Gene (0kb buffer model)	3.820E-03**	1.64E-03	3.296E-03*	1.63E-03	5.459E-03***	1.79E-03	4.11E-3 [2.20E-3,6.00E-3]	4.96%
Gene (5kb buffer model)	2.865E-03*	1.68E-03	1.54E-03	1.62E-03	4.184E-03**	1.86E-03	2.74E-3 [0.80E-3,4.70E-3]	3.26%
Gene (10kb buffer model)	1.72E-03	1.60E-03	8.89E-04	1.56E-03	2.933E-03*	1.78E-03	1.76E-3 [-0.10E-3,3.60E-3]	2.16%
Gene (25kb buffer model)	1.26E-03	1.57E-03	1.07E-03	1.58E-03	2.773E-03*	1.76E-03	1.60E-3 [-0.20E-3,3.50E-3]	1.92%
Gene (35kb buffer model)	1.34E-03	1.58E-03	1.31E-03	1.60E-03	2.995E-03*	1.77E-03	1.81E-3 [<-0.01E-3,3.70E-3]	2.16%
Gene (50kb buffer model)	3.117E-03*	1.61E-03	2.600E-03*	1.59E-03	4.947E-03**	1.78E-03	3.50E-3 [1.60E-3,5.30E-3]	4.17%
ROI - Buffer								
Buffer 0kb	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Buffer 5kb	3.253E-03*	1.78E-03	5.822E-03***	1.88E-03	3.431E-03*	1.86E-03	4.13E-3 [2.00E-3,6.20E-3]	4.95%
Buffer 10kb	7.880E-03***	2.08E-03	8.964E-03***	2.10E-03	7.717E-03***	2.13E-03	8.19E-3 [5.80E-3,1.06E-2]	9.84%
Buffer 25kb	1.101E-02***	2.35E-03	9.625E-03***	2.27E-03	1.016E-02***	2.39E-03	1.02E-2 [7.60E-3,1.29E-2]	12.36%
Buffer 35kb	1.135E-02***	2.44E-03	9.542E-03***	2.35E-03	1.048E-02***	2.48E-03	1.04E-2 [7.70E-3,1.32E-2]	12.50%
Buffer 50kb	3.141E-02***	5.29E-03	3.303E-02***	5.28E-03	2.743E-02***	5.32E-03	3.06E-2 [2.46E-2,3.66E-2]	36.47%
ROI - Other genomewide variants								
All Other Variants (0kb buffer model)	7.145E-02**	7.82E-03	7.586E-02***	7.76E-03	8.853E-02***	8.04E-03	7.84E-2 [6.95E-2,8.73E-2]	94.92%
All Other Variants (5kb buffer model)	6.923E-02***	7.85E-03	7.235E-02***	7.76E-03	8.658E-02***	8.06E-03	7.58E-2 [6.69E-2,8.48E-2]	91.44%
All Other Variants (10kb buffer model)	6.607E-02***	7.80E-03	7.042E-02***	7.73E-03	8.408E-02***	8.03E-03	7.33E-2 [6.44E-2,8.22E-2]	88.00%
All Other Variants (25kb buffer model)	6.359E-02***	7.75E-03	6.933E-02***	7.70E-03	8.184E-02***	7.99E-03	7.14E-2 [6.25E-2,8.02E-2]	85.71%
All Other Variants (35kb buffer model)	6.319E-02***	7.73E-03	6.898E-02***	7.68E-03	8.105E-02***	7.97E-03	7.09E-2 [6.20E-2,7.97E-2]	85.22%
All Other Variants (50kb buffer model)	4.216E-02***	7.40E-03	4.524E-02***	7.34E-03	6.242E-02***	7.72E-03	4.96E-2 [4.11E-2,5.80E-2]	59.12%
Total								
Total heritability (0kb buffer model)	7.53E-02	7.86E-03	7.92E-02	7.79E-03	9.40E-02	8.08E-03	8.26E-2 [7.36E-2,9.15E-2]	N/A
Total heritability (5kb buffer model)	7.53E-02	7.85E-03	7.97E-02	7.78E-03	9.42E-02	8.08E-03	8.29E-2 [7.39E-2,9.18E-2]	N/A
Total heritability (10kb buffer model)	7.57E-02	7.84E-03	8.03E-02	7.78E-03	9.47E-02	8.07E-03	8.33E-2 [7.44E-2,9.23E-2]	N/A
Total heritability (25kb buffer model)	7.59E-02	7.84E-03	8.00E-02	7.77E-03	9.48E-02	8.07E-03	8.33E-2 [7.44E-2,9.22E-2]	N/A
Total heritability (35kb buffer model)	7.59E-02	7.84E-03	7.98E-02	7.78E-03	9.45E-02	8.07E-03	8.32E-2 [7.43E-2,9.21E-2]	N/A
Total heritability (50kb buffer model)	7.67E-02	7.85E-03	8.09E-02	7.79E-03	9.48E-02	8.09E-03	8.39E-2 [7.49E-2,9.28E-2]	N/A

Table shows the estimated heritability for each fold and the meta-heritability estimated across folds. Note that components are labelled according to the observed effects used across the models with varied buffer lengths.

Table 3. Calculated Enrichment Values of Model Components

Model Component	Number of SNPs	Observed h^2_{meta}	Expected h^2_{meta}	Enrichment
Gene (0kB buffer model)	81453	0.0041	0.0015	2.82
Gene (5kB buffer model)	81453	0.0027	0.0015	1.88
Gene (10kB buffer model)	81453	0.0018	0.0015	1.21
Gene (25kB buffer model)	81453	0.0016	0.0015	1.10
Gene (35kB buffer model)	81453	0.0018	0.0015	1.24
Gene (50kB buffer model)	81453	0.0035	0.0015	2.41
Buffer 0kB	N/A	N/A	N/A	N/A
Buffer 5kB	10815	0.0041	0.0002	21.37
Buffer 10kB	21288	0.0082	0.0004	21.53
Buffer 25kB	53341	0.0102	0.0010	10.70
Buffer 35kB	74436	0.0104	0.0013	7.82
Buffer 50kB	841092	0.0306	0.0150	2.04
All Other Variants (0kB buffer model)	4575485	0.0784	0.0817	0.96
All Other Variants (5kB buffer model)	4564670	0.0758	0.0816	0.93
All Other Variants (10kB buffer model)	4554197	0.0733	0.0814	0.90
All Other Variants (25kB buffer model)	4522144	0.0714	0.0808	0.88
All Other Variants (35kB buffer model)	4501049	0.0709	0.0804	0.88
All Other Variants (50kB buffer model)	3734393	0.0496	0.0667	0.74
Total	4656938	0.0832		

Note: E >= 1.96 consitutes statistically significant enrichment at a 95% CI

Table 4. Polygenic Model 1 Standardized Results for Cigarettes Per Day				
	β_1 Estimate	**S.E.**	**Estimate/S.E.**	**P-value**
PC1	-0.040	0.024	-1.656	0.098
PC2	0.032	0.024	1.319	0.187
PC3	-0.058	0.024	-2.374	0.018
PC4	0.012	0.024	-0.499	0.618
PC5	0.006	0.024	0.260	0.795
PC6	0.003	0.024	0.133	0.894
All Variants Score	0.174	0.024	7.258	<0.001
	R-squared	**S.E.**	**Estimate/S.E.**	**P-value**
Full Model	0.035	0.009	3.92	0.000

Table 5. Strength of Associations and Variance Explained by PRS of All Variants					
Observed Phenotype	**β_1 Estimate**	**Upper 2.5%**	**Lower 2.5%**	**S.E.**	**Partial R_2**
CPD	0.174	0.221	0.127	0.024	0.0303
Frequency	0.122	0.171	0.073	0.025	0.0148
Age of First Use	-0.039	0.011	-0.089	0.025	~0
Age of Initiation	-0.079	-0.028	-0.131	0.026	0.0063
Fagerstrom Index	0.164	0.213	0.115	0.025	0.0268

Table 6. Polygenic Model 2 Standardized Results for Cigarettes Per Day				
	β_1 Estimate	**S.E.**	**Estimate/S.E.**	**P-value**
PC1	-0.039	0.024	-1.612	0.107
PC2	0.034	0.024	1.424	0.154
PC3	-0.057	0.024	-2.337	0.019
PC4	-0.013	0.024	-0.552	0.581
PC5	0.008	0.024	0.316	0.752
PC6	0.004	0.024	0.152	0.879
Gene Score	0.079	0.027	2.884	0.004
Buffer Score	-0.023	0.028	-0.840	0.401
Others Score	0.164	0.025	6.669	0.000
	R-squared	**S.E.**	**Estimate/S.E.**	**P-value**
Full Model	0.038	0.009	4.113	0.000

Table 7. Regression Coefficients and Variance Explained by Partitioned Risk Scores

Observed Phenotype	β_1 Estimate	Upper 2.5%	Lower 2.5%	Number of SNPs	Observed R_2	Expected R_2	Enrichment	Distribution P-value
CPD (Gene)	0.079	0.132	0.025	81453	0.0062	0.0011	5.88	0.020
CPD (Buffer)	-0.023	0.031	-0.078	21288	~0	0.0007	0.00	0.380
CPD (All Others)	0.164	0.212	0.116	4554197	0.0269	0.0300	0.90	0.610
Freq. (Gene)	0.068	0.124	0.013	81453	0.0046	0.0011	4.22	0.031
Freq. (Buffer)	-0.031	0.025	-0.087	21288	~0	0.0007	0.00	0.125
Freq. (All Others)	0.116	0.166	0.066	4554197	0.0134	0.0150	0.90	0.451
Age of First Use (Gene)	-0.045	0.011	-0.101	81453	~0	0.0013	0.00	0.105
Age of First Use (Buffer)	-0.007	0.050	-0.063	21288	~0	0.0007	0.00	0.391
Age of First Use (All Others)	-0.029	0.022	-0.080	4554197	~0	-0.0388	0.00	0.366
Age of Initiation (Gene)	-0.009	0.049	-0.066	81453	~0	0.0013	0.00	0.418
Age of Initiation (Buffer)	-0.032	0.026	-0.090	21288	~0	0.0008	0.00	0.865
Age of Initiation (All Others)	-0.071	-0.018	-0.123	4554197	0.0050	0.0069	0.73	0.383
Fagerstrom (Gene)	0.059	0.115	0.004	81453	0.0035	0.0012	3.00	0.059
Fagerstrom (Buffer)	0.015	0.071	-0.041	21288	~0	0.0008	0.00	0.302
Fagerstrom (All Others)	0.148	0.199	0.098	4554197	0.0220	0.0265	0.83	0.305

*See Methods section for full description of dependent variables

Collection of Figures

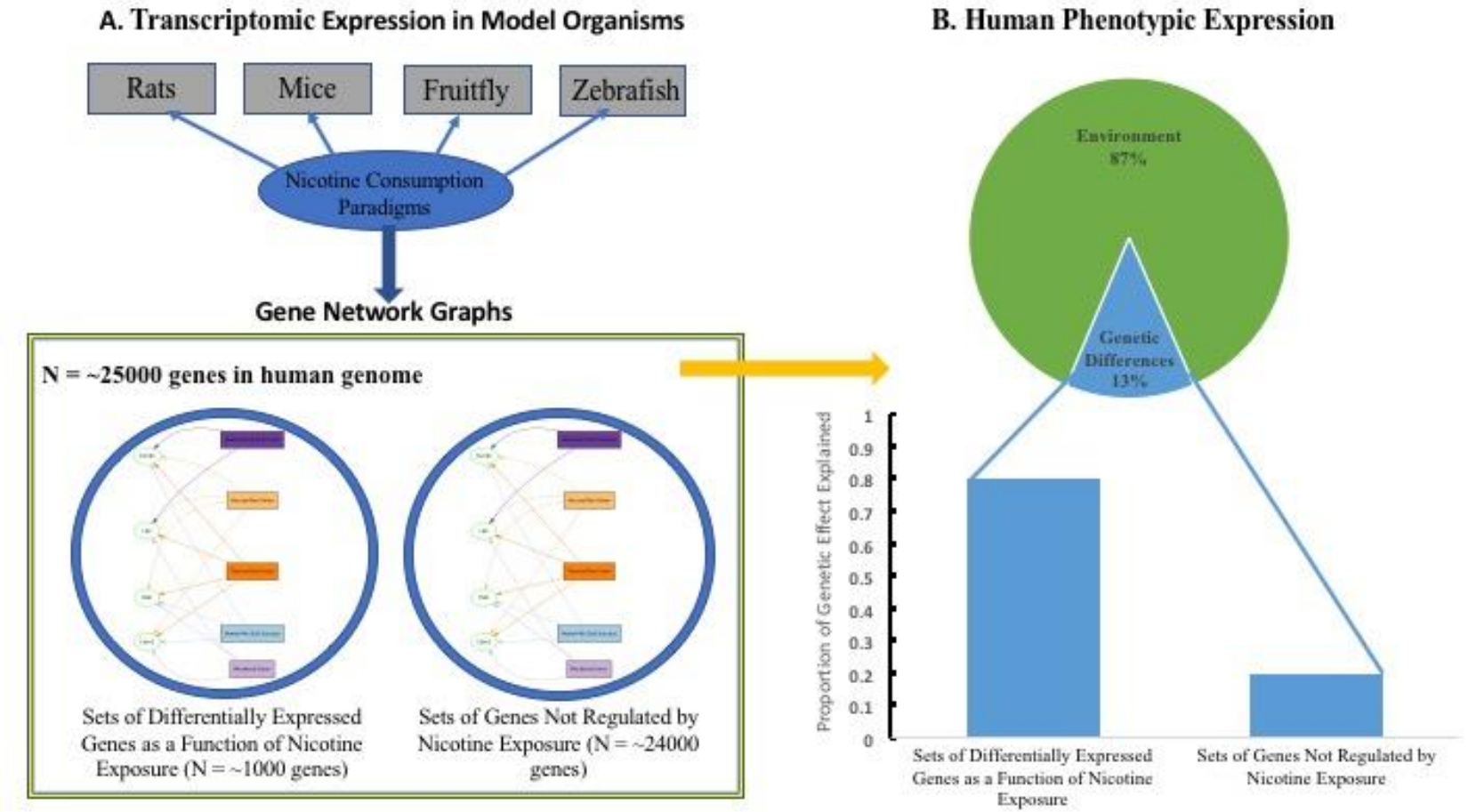

Figure 1. Theoretical integrative genomics approach to assessing translatability of expression data from model organisms to capturing genetic variance of nicotine consumption in human populations.

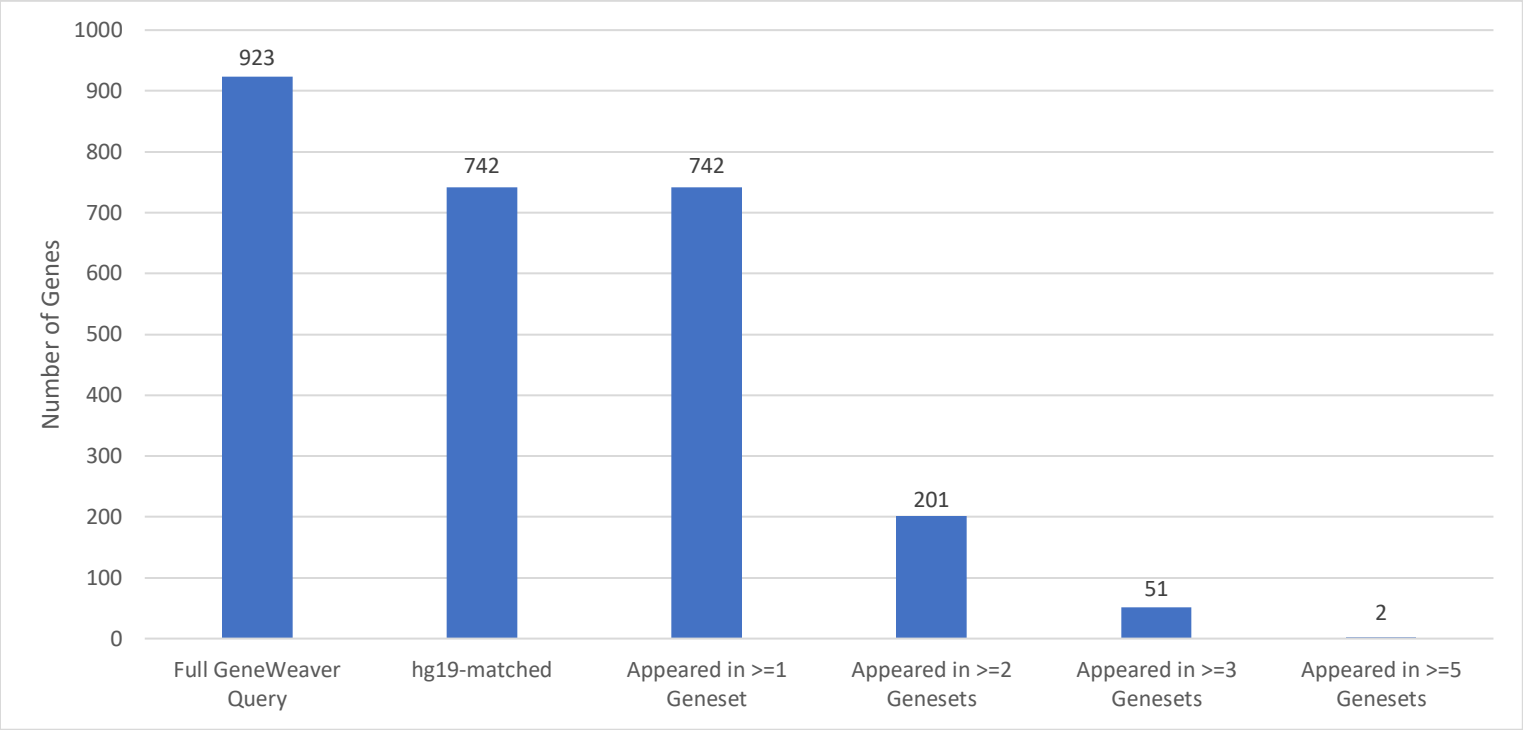

Figure 2. Distribution of genes identified via GeneWeaver query.

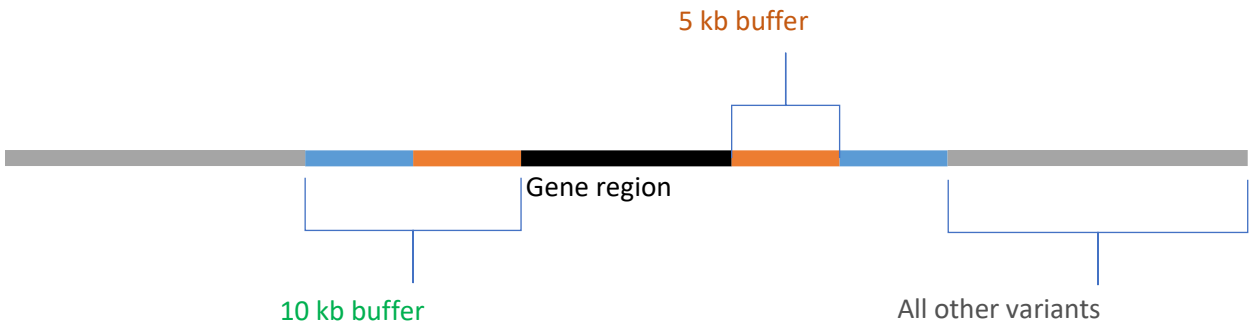

Figure 3. Visualization of each region-of-interest used as a model component in statistical analyses.

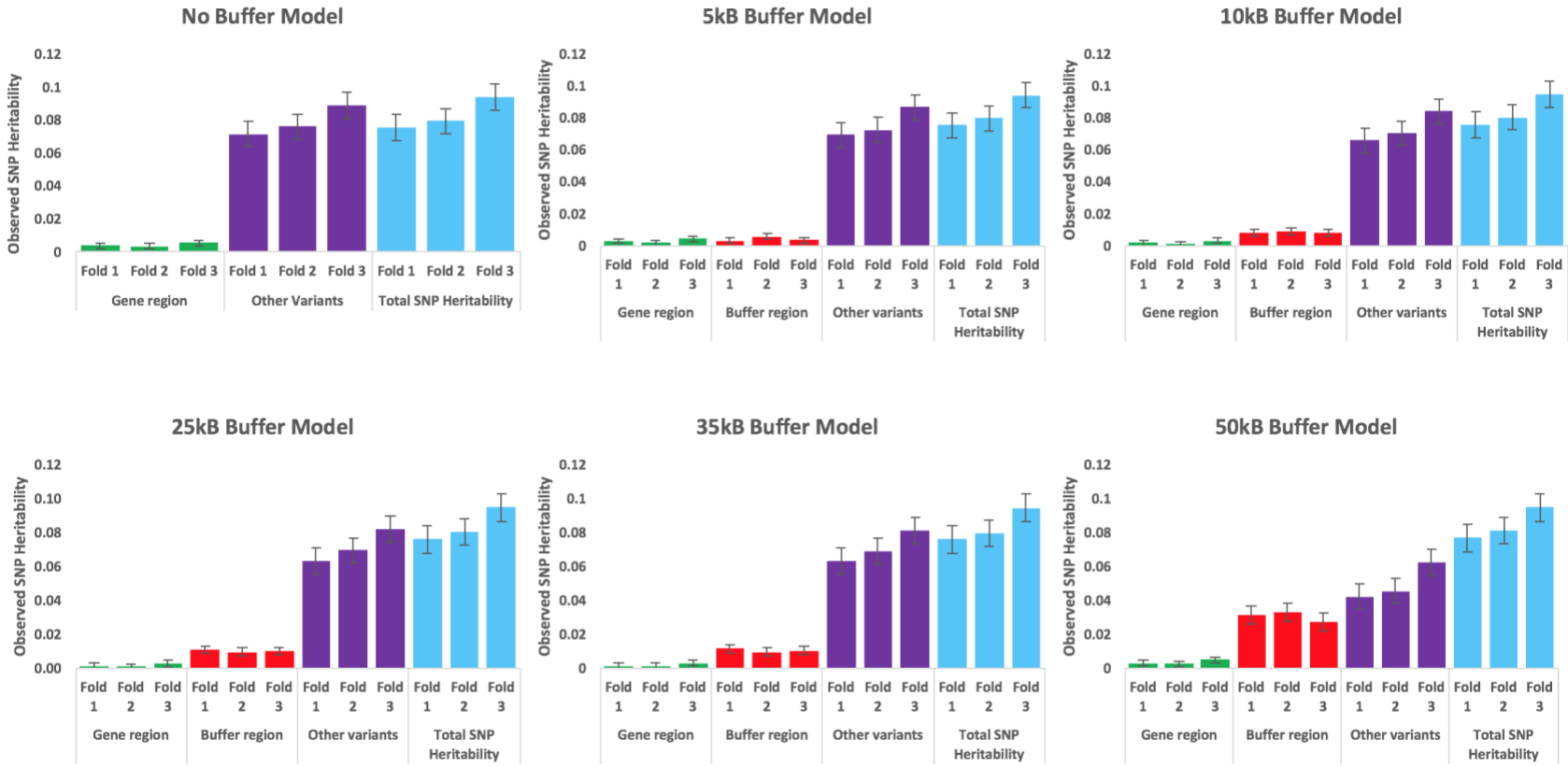

Figure 4. Visualization of partial R_2 accounted for by regions of interest across various multivariate models.

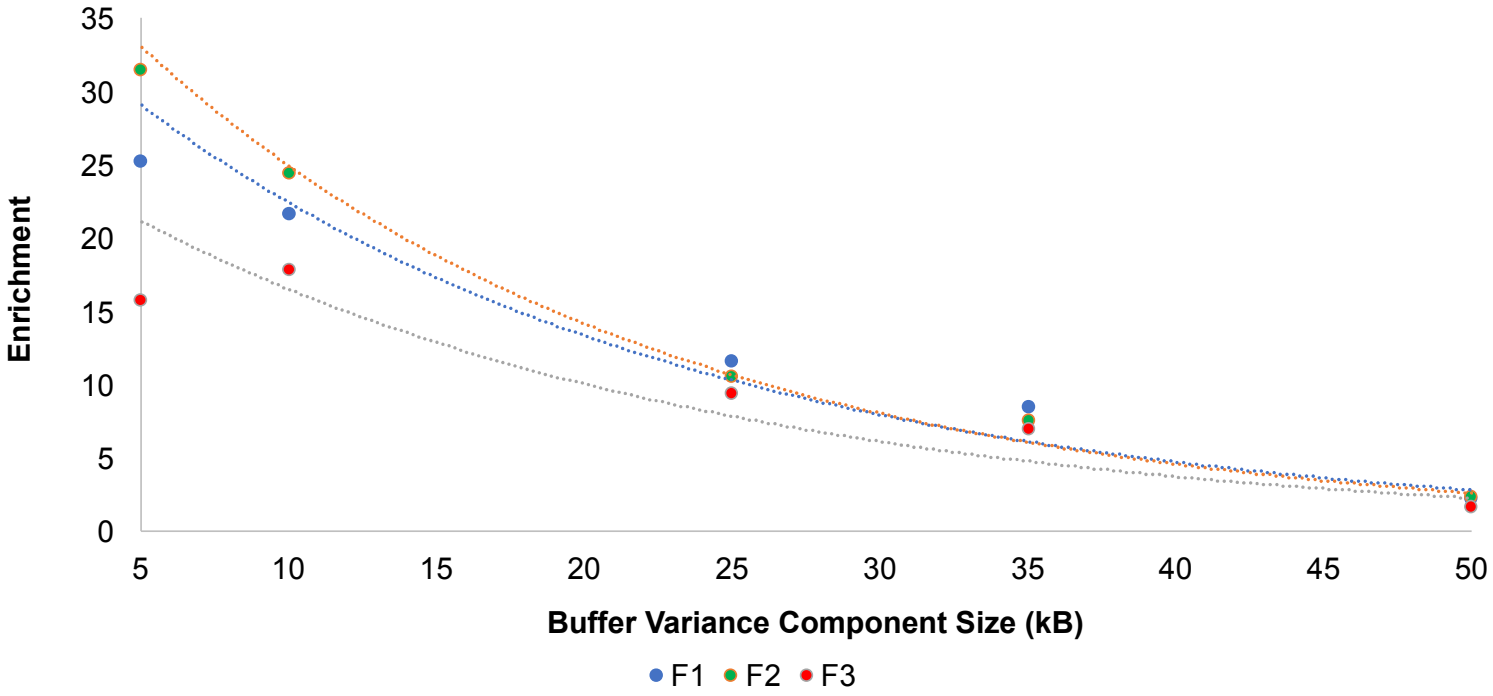

Figure 5. Enrichment decay of flanking buffer region seen with increasing buffer size.

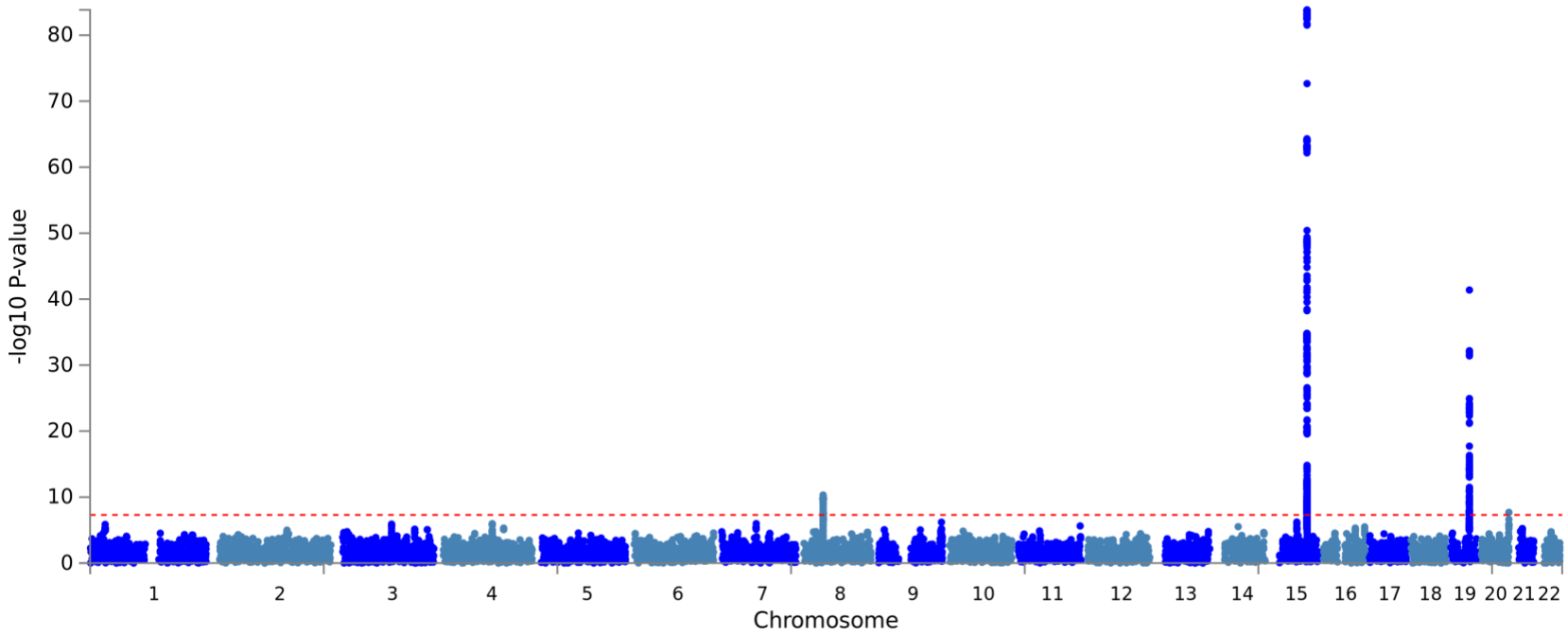

Figure 6. Manhattan plot of UK Biobank MLMA for nicotine/tobacco consumption.

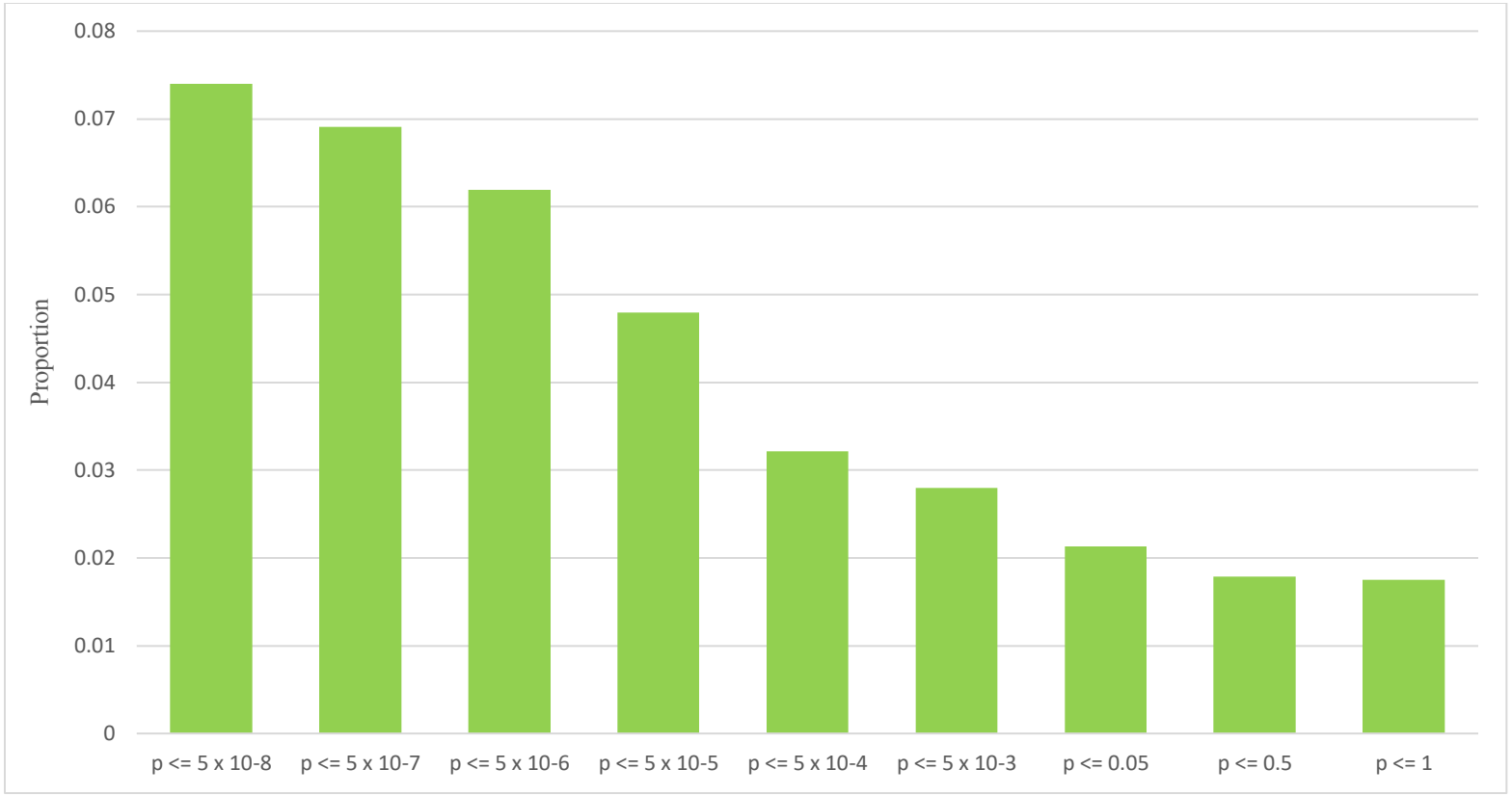

Figure 7. Proportion of SNPs across successive bins of UK Biobank GWAS p-value distribution that are found in the prioritized subset of genes obtained from animal model expression data.