Addiction-associated genetic variants implicate brain cell type- and region-specific cis-regulatory elements in addiction neurobiology

Chaitanya Srinivasan[1*], BaDoi N. Phan[1,2*], Alyssa J. Lawler[3], Easwaran Ramamurthy[1], Michael Kleyman[1], Ashley R. Brown[1], Irene M. Kaplow[1,4], Morgan E. Wirthlin[1,4], Andreas R. Pfenning [1,3,4,‡]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University, 15213
[2]Medical Scientist Training Program, School of Medicine, University of Pittsburgh, 15213
[3]Department of Biological Sciences, Mellon College of Science, Carnegie Mellon University, 15213
[4]Neuroscience Institute, Carnegie Mellon University, 15213
[*]CS and BNP contributed equally to the work
[‡]Direct correspondence to apfenning@cmu.edu

Number of pages: 54 pages of text, 13 pages of figures

Number of figures:
- 5 main figures
- 6 supplemental figures
- 1 supplemental table

Number of words
- Abstract: 244 words
- Introduction: 559 words
- Discussion: 1347 words

CONFLICT OF INTEREST:
AJL, ER, and ARP are inventors of the cSNAIL patent. Other authors do not declare any conflict of interest.

ACKNOWLEDGEMENT

We would like to thank members of the Eric Yttri lab at Carnegie Mellon University for providing *Drd1-cre* and *Adora2a-cre* mice for cell type-specific ATAC-seq experiments.

Author contributions: Conceptualization: ARP, BNP, CS; ATAC-seq data processing: AJL, ER, IMK, BNP; GWAS enrichment investigation: BNP, CS, ER, MK; Machine learning models: AJL, BNP, ER, IMK; bulk tissue ATAC-seq: AJL, ARB, MEW; cSNAIL ATAC-seq: AJL, ARB; writing (original draft): BNP, CS; review and editing: AJL, ARP, BNP, CS, ER, IMK; funding acquisition, resources, & supervision: ARP;

Funding: National Institute of General Medical Sciences training grant T32GM008208 (BNP), Sloan Foundation Fellowship (ARP), National Institute on Drug Abuse Avenir Award 1DP1DA046585 (ARP), National Science Foundation Graduate Student Research Fellowship DGE1745016 (AJL), Carnegie Mellon Brainhub Presidential Fellowship (ER), Carnegie Mellon Computational Biology Department Lane Postdoctoral Fellowship (IMK)

ABSTRACT

Recent large genome-wide association studies (GWAS) have identified multiple confident risk loci linked to addiction-associated behavioral traits. Genetic variants linked to addiction-associated traits lie largely in non-coding regions of the genome, likely disrupting cis-regulatory element (CRE) function. CREs tend to be highly cell type-specific and may contribute to the functional development of the neural circuits underlying addiction. Yet, a systematic approach for predicting the impact of risk variants on the CREs of specific cell populations is lacking. To dissect the cell types and brain regions underlying addiction-associated traits, we applied LD score regression to compare GWAS to genomic regions collected from human and mouse assays for open chromatin, which is associated with CRE activity. We found enrichment of addiction-associated variants in putative regulatory elements marked by open chromatin in neuronal (NeuN+) nuclei collected from multiple prefrontal cortical areas and striatal regions known to play major roles in reward and addiction. To further dissect the cell type-specific basis of addiction-associated traits, we also identified enrichments in human orthologs of open chromatin regions of mouse neuron subtypes: cortical excitatory, PV, D1, and D2. Lastly, we developed machine learning models from mouse cell type-specific regions of open chromatin to further dissect human NeuN+ open chromatin regions into cortical excitatory or striatal D1 and D2 neurons and predict the functional impact of addiction-associated genetic variants. Our results suggest that different neuron subtypes within the reward system play distinct roles in the variety of traits that contribute to addiction.

Significance Statement:

Our study on cell types and brain regions contributing to heritability of addiction-associated traits suggests that the conserved non-coding regions within cortical excitatory and striatal medium

68 spiny neurons contribute to genetic predisposition for nicotine, alcohol, and cannabis use
69 behaviors. This computational framework can flexibly integrate epigenomic data across species to
70 screen for putative causal variants in a cell type- and tissue-specific manner across numerous
71 complex traits.

72 **INTRODUCTION**

73 Substance use disorders (SUD) have increased in prevalence over the last three decades,
74 with an estimated 100 million cases worldwide(GBD 2016 Alcohol and Drug Use Collaborators,
75 2018; Eddie et al., 2019). Pharmacological interventions are limited in their ability to cure
76 addiction due to physiological and logistical barriers(Pullen and Oser, 2014; Pear et al., 2019). As
77 the societal epidemic of substance use grows, there is a greater need to understand the neurobiology
78 of substance use behaviors and addiction.

79 The reward circuits co-opted in addiction are highly conserved across primates and
80 rodents(Scaplen and Kaun, 2016). It is generally accepted that addictive substances promote
81 impulsive and compulsive behavior by activating the mesolimbic dopamine system, in which
82 dopaminergic inputs from the ventral tegmental area project to medium spiny neurons (MSN) of
83 the nucleus accumbens (NAc) (Koob and Volkow, 2010). Furthermore, glutamatergic inputs to the
84 NAc from the amygdala, frontal cortex, and hippocampus contribute to motivational action
85 through the extrapyramidal motor system(Koob and Volkow, 2010). Subsequently, the NAc sends
86 outputs to different nuclei of the ventral pallidum critical for processing and modulating substance
87 reward signal(Koob and Volkow, 2010). The dorsal striatum (STR) is also hypothesized to be
88 recruited during the development of compulsive substance-seeking(Koob and Volkow, 2010).
89 Additionally, the dorsal striatum and prefrontal cortical regions(Goldstein and Volkow, 2011)
90 regulate reward and addiction-related phenotypes. The findings emphasize the interplay of brain
91 regions and the cellular components involved in substance use behavior.

92 Increasing evidence reveals strong genetic links to substance use risk(Pasman et al., 2018;
93 Erzurumluoglu et al., 2019; Karlsson Linnér et al., 2019; Liu et al., 2019b) and SUD(Kendler and
94 Prescott, 1998a, 1998b; Dick, 2016; Waaktaar et al., 2018). Genome-wide association studies

95 (GWAS) report that genetic risk for substance use shares underlying architecture with other
96 neuropsychiatric disorders(Pasman et al., 2018; Liu et al., 2019b), of which risk variants tend to
97 lie in non-coding and functional regions of the human genome(Jensen, 2016). These single
98 nucleotide polymorphisms (SNPs) can disrupt transcription factor binding in cis-regulatory
99 elements (CREs) with varying impact on gene regulation and downstream neural circuitry. Many
100 CREs have tissue- and cell type-specific activity(Roadmap Epigenomics Consortium et al., 2015),
101 suggesting that cell types and tissues underlying addiction may be uniquely targeted by genetic
102 variants at these CREs. GWAS for nicotine-, alcohol-(Liu et al., 2019b), and cannabis-use
103 traits(Pasman et al., 2018) have identified multiple confident risk loci and SNPs linked to
104 addiction-associated phenotypes with brain-specificity, yet their effects on the CREs of specific
105 brain regions and cell types involved in addiction pathophysiology are an open area of inquiry.

106 GWAS have found that SNP enrichment within functional non-coding regions contribute
107 disproportionately to heritability of complex traits due to polygenicity(Finucane et al., 2015).
108 Linkage disequilibrium (LD) of significant SNPs complicates the identification of causal variants
109 contributing to genetic risk(Bush and Moore, 2012). The functional consequences of risk SNPs in
110 CRE sequences cannot be reliably inferred from DNA sequences alone(Shlyueva et al., 2014).
111 Recent developments in epigenomic assays(Buenrostro et al., 2013; Mo et al., 2015; Tak and
112 Farnham, 2015) and machine learning(Ghandi et al., 2014; Zhou and Troyanskaya, 2015; Kelley
113 et al., 2016, 2018; Lee, 2016) can predict cell types affected by addiction-associated genetic
114 variation to propose cell type-specific hypotheses on the pathogenesis of addiction.

115 Here, we implement a framework that identifies regions and cell types that are affected by
116 genetic risk variants for addiction-associated traits. We intersect these GWAS variants across
117 human and mouse datasets from bulk tissue and cell type-specific open chromatin assays to

118 identify region- and cell type-specific CREs that may be impacted by genetic variation associated
119 with addiction-related traits. We then apply machine learning models to predict open chromatin
120 activity in key neuronal subtypes within reward circuits and refine our GWAS enrichments in these
121 subtypes. Further, we apply these models to screen for putative causal SNPs within dense loci
122 reported in GWAS for addiction-associated traits. This pipeline, to our knowledge, describes the
123 first integrative analyses across species, regions and cell types to screen for candidate causal
124 addiction-associated genetic risk variants in dense loci with numerous significant SNPs in LD.

125 **RESULTS**

126 **Genetic risk for substance use traits is associated with the neuronal epigenomes of reward**
127 **areas**

128 Recent well-powered GWAS have characterized the genetic loci associated with seven
129 addiction-associated traits, revealing several candidate loci encoding addiction genetic
130 risk(Pasman et al., 2018; Karlsson Linnér et al., 2019; Liu et al., 2019b). These GWAS identify
131 genetic variants associated with reward, risk tolerance, and various substance use behaviors to
132 provide a means of studying genetic variation associated with addiction. We found that 72-98% of
133 addiction-associated genetic variants map to noncoding intronic or intergenic regions of the
134 genome **(Figure 1A)**. The proportion of intronic risk variants (47%-85%) overrepresented the
135 proportion of intronic variants on the reference genotype panel (odds ratio, $OR_{AgeOfInit}$ =2.3,
136 $OR_{Cannabis}$ = 2.3, $OR_{CigsPerDay}$ = 1.4, $OR_{DrinksPerWeek}$ = 1.6, $OR_{RiskyBehavior}$ = 1.4, $OR_{SmokCess}$ = 1.8,
137 $OR_{SmokInit}$ =1.3, Fisher's Exact $P_{Bonferroni} < 8 \times 10^{-79}$). Furthermore, these seven traits have shared
138 and distinct genetic architecture estimated using genetic correlation of risk alleles (r_g,
139 **Supplemental Figure 1A**). Although common genetic variants are shared between addiction-
140 associated traits on a genome-wide scale, the reported significant loci are often unique to a
141 particular trait and are densely packed with SNPs in high LD (**Supplemental Figure 1B**).

142 We investigated whether genetic variants of addiction-associated GWAS cluster at putative
143 cis-regulatory elements (CREs) of the brain using a partitioned heritability linkage disequilibrium
144 score (LDSC) regression approach(Bulik-Sullivan et al., 2015b; Finucane et al., 2018) on open
145 chromatin regions (OCRs) of sorted neuronal (NeuN+) and glial (NeuN-) nuclei across 14 brain
146 regions(Fullard et al., 2018) **(Figure 1B)**. We found that genetic variants associated with having
147 ever regularly smoked (`SmokingInitiation`), being a former versus current smoker

148 (SmokingCessation), the number of alcoholic drinks per week (DrinksPerWeek), and
149 lifetime cannabis use (Cannabis) significantly enriched in NeuN+ OCRs of brain regions
150 known and speculated to contribute to reward and addiction(Volkow and Morales, 2015) (FDR <
151 0.05). In particular, we found that genetic variants associated with SmokingInitiation and
152 Cannabis shared enrichment in NeuN+ prefrontal cortical OCRs (from orbitofrontal cortex and
153 dorsolateral prefrontal cortex) while those associated with SmokingCessation and
154 DrinksPerWeek shared enrichment in NeuN+ striatal OCRs (both putamen and NAc). The
155 enrichments of NeuN+ OCRs indicate that genetic variation in epigenomes of neuronal
156 populations from frontal cortex and striatum contribute to addiction liability. The difference in
157 NeuN+ enrichments between regions across addiction-associated traits can likely be explained by
158 the difference in proportions of neuronal subtypes of each area, so we sought to dissect the different
159 neuronal subtypes contributing to these enrichments.

160 Broad marker-gene based labeling approaches, such as using NeuN to label neurons, do
161 not capture the rich diversity of neuronal subtypes; bulk NeuN+ open chromatin signal represents
162 an average signal from heterogeneous neuronal subtypes, each with distinct epigenomic
163 landscapes, gene regulation, network connectivity. Hence, NeuN-labeled open chromatin profiles
164 likely do not capture OCRs unique to less populous neuronal subtypes. The difference in
165 proportions of neuronal subtypes between brain regions may also contribute to brain region-
166 specific NeuN+ OCR enrichment for GWAS variants of addiction-associated traits. We therefore
167 applied LDSC regression GWAS enrichment on open chromatin profiles from human postmortem
168 occipital cortex cell types via single-cell transposase hypersensitivity sequencing (scTHS-seq)
169 (Lake et al., 2018) **(Figure 1C)**. We found that addiction-associated genetic variants largely
170 enriched in both excitatory and inhibitory neuronal OCRs. We found enrichment of genetic

171 variants associated with age of smoking initiation (AgeOfInitiation) and
172 SmokingCessation in OCRs of cortical excitatory neurons. We found no enrichment of
173 genetic variants associated with average number of cigarettes smoked per day
174 (CigarettesPerDay) for OCRs of occipital cortex cell types. Genetic variants associated with
175 SmokingInitiation, which enriched in astrocyte, endothelial, inhibitory, and
176 oligodendrocyte precursor cell OCRs, shared enrichment in NeuN- OCRs of mediodorsal thalamus
177 **(Figure 1B)**. Interestingly, genetic variants associated with SmokingCessation, which had
178 enriched for striatal NeuN+ OCRs, enriched only for OCRs of cortical excitatory neurons and not
179 cortical inhibitory neurons. Sorted bulk ATAC-seq only showed enrichment of
180 SmokingCessation associated genetic variants in OCRs of NeuN+ striatal regions, which are
181 largely composed of inhibitory MSNs. Single-cell epigenomics of human postmortem brain can
182 further dissect the genetic risk for substance-use traits into neuronal subtypes that otherwise would
183 not be parsed with bulk tissue assays.

184 We confirmed that our pipeline for LDSC regression on NeuN-sorted OCRs from 14 brain
185 regions is able to reproduce the GWAS enrichments published by Fullard *et al.* While our approach
186 uses OCRs from reproducible ATAC-seq peaks rather than differentially accessible peaks, we
187 found consistent enrichments of genetic variants associated with schizophrenia risk
188 (Schizophrenia), highest level of educational attainment (EduAttain), and habitual sleep
189 duration (SleepDuration) **(Supplemental Figure 2B)**. We did not find enrichment in brain
190 OCRs of genetic variants identified in several low-powered GWAS (cocaine dependence
191 (CocaineDep)(Cabana-Domínguez et al., 2019), opioid dependence (OpioidDep) (Cheng et
192 al., 2018), and obsessive-compulsive disorder (OCD)(International Obsessive Compulsive
193 Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics

194 Association Studies (OCGAS), 2018), each of which had included fewer than 5000 individuals
195 with the trait **(Supplemental Figure 2A)**. In addition, we found no enrichments in brain OCR for
196 several well-powered studies of traits related to addiction behaviors, multi-site chronic pain
197 (`ChronicPain`)(Johnston et al., 2019) and cups of coffee per day (`CoffeePerDay`)(Coffee
198 and Caffeine Genetics Consortium et al., 2015) and anthropometric traits coronary artery disease
199 (`CAD`) (Howson et al., 2017), bone mineral density (`BMD`) (Kemp et al., 2017), and lean body mass
200 (`LBM`) (Zillikens et al., 2017) **(Supplemental Figure 2B, C)**. Lastly, we validated that human
201 OCRs from non-brain tissues would not enrich for risk variants associated with brain traits. We
202 gathered publicly available OCRs from stomach ATAC-seq, adipocyte ATAC-seq, preadipocyte
203 ATAC-seq, liver DNase-seq, and lung DNase-seq profiles(ENCODE Project Consortium, 2012;
204 Thurman et al., 2012; Davis et al., 2018; Cannon et al., 2019) **(Supplemental Figure 4D)** and
205 performed LDSC regression on the total 18 GWAS from above. To our expectation, we did not
206 find enrichments of stomach, liver, or lung OCRs for genetic variants associated with brain-related
207 traits. We did find enrichment of `BMD` in lung OCRs, a connection previously recognized(Lee et
208 al., 2016; Kim et al., 2019; Zeng et al., 2019). The secondary GWAS enrichments in other traits
209 and foregrounds demonstrate two trends: a GWAS trait would enrich if the GWAS was properly
210 powered to detect genetic risk variants, and the foreground regions are from cell types or tissue of
211 that trait's potential etiological origin.
212

213 **Mouse-human conserved cell type-specific open chromatin enrich for addiction risk loci**

214 In order to resolve the different neuronal subtypes that comprise the addiction-associated
215 enrichments in Fullard *et al.* and Lake *et al.* OCRs **(Figure 1, Supplemental Figure 2)**, we
216 performed targeted epigenomics in mouse to isolate various neuronal subtypes from frontal cortex

217 (CTX), caudoputamen (CPU), and the nucleus accumbens (NAc), which are key brain regions of
218 the reward circuit. We isolated cell type-specific nuclei for ATAC-seq using a modified version
219 of the INTACT approach(Mo et al., 2015), which we call *cre*-specific nuclei anchored independent
220 labeling (cSNAIL). cSNAIL-INTACT uses the AAV-PHP.eB virus to isolate nuclei marked by
221 *Pvalb*, *Sst*, *Drd1*, and *Adora2a cre*-lines without crossing with a *Sun1-Gfp* transgenic mouse
222 **(Figure 2A)**. We show that cell type-targeting provided markedly distinct genome-wide ATAC-
223 seq profiles compared to bulk tissue ATAC-seq alone (**Supplemental Figure 3A**). cSNAIL
224 ATAC-seq specifically captured nuclei with increased accessibility around the marker gene that
225 was driving *Cre* recombinase expression (**Supplemental Figure 3B**). Accessibility around
226 cSNAIL ATAC-seq transcription start sites (TSS) strongly correlated with matched pseudobulk
227 gene expression in the same cell type and tissue (**Methods**, both Pearson and Spearman correlation
228 $P_{bonf} < 2 \times 10^{-16}$, **Supplemental Figure 3C,D**). We applied the HALPER approach(Zhang et al.,
229 2020) to reliably map ~70% of mouse OCRs to their human orthologs in hg38 **(Methods)** to run
230 LDSC regression GWAS enrichments.

231 Our GWAS enrichments of human orthologs from mouse OCRs of various neuronal
232 subtypes and bulk tissue **(Figure 2B)** show that genetic variants associated with
233 `SmokingInitiation` and `Cannabis` shared enrichment in human orthologs of mouse
234 cortical PV and EXC neuron OCRs from both Mo *et al.* and this study (FDR < 0.05). Cannabis
235 associated genetic variants further enriched in CTX bulk tissue OCRs, which could be attributed
236 to signal from cortical EXC and PV neuron populations. Cortical PV neuron OCRs further
237 enriched with genetic variants associated with `DrinksPerWeek`. `SmokingCessation`
238 associated genetic variants distinctly enriched in cortical VIP neuron OCRs.

239 Within cell types from CPU and NAc, we found enrichment of genetic variants associated
240 with all measured addiction-associated traits in CPU and NAc D2 MSN OCRs. Genetic variants
241 associated with all measured traits excluding `SmokingInitiation` and `RiskyBehavior` all
242 enriched in CPU and NAc D1 MSN OCRs. CPU D1 MSN OCRs were enriched with genetic
243 variants associated with all measured traits excluding `RiskyBehavior`. We found that CPU
244 bulk tissue OCRs were enriched with genetic variants associated with all measured addiction-
245 associated traits excluding `AgeOfInitiation` and `RiskyBehavior`. Distinctly, CPU PV+
246 and SST+ neuron OCRs enriched with genetic variants associated with `Cannabis`.

247 Corresponding to our analysis of human brain OCRs, we also confirmed the specificity of
248 mouse-human orthologous CRE enrichments for genetic variants of addiction-related, brain-
249 related, and non-brain related traits **(Supplemental Figure 4)**. We found enrichments of genetic
250 variants associated with `ChronicPain` in cortical PV neuron OCRs from both Mo *et al.* and this
251 study **(Supplemental Figure 4A).** Within striatal cell types, we found that CPU D2 and NAc D1
252 MSN OCRs were enriched for genetic variants associated with `ChronicPain`. In contrast, CPU
253 D1 and NAc D2 MSN OCRs were enriched for genetic variants associated with `OpioidDep`.
254 Genetic variants associated with `OpioidDep` also enriched in CPU D1 MSN and CPU PV OCRs.
255 `Schizophrenia`, `EduAttain`, and `SleepDuration` associated genetic variants all
256 enriched in OCRs of all measured cell types **(Supplemental Figure 4B)**. None of these human
257 orthologs of mouse neuronal OCRs enriched for genetic variants associated with non-brain-related
258 traits `BMD`, `CAD`, and `LBM` **(Supplemental Figure 4C)**. We validated that our approach to map
259 OCRs from mouse to human did not bias enrichment to brain traits by performing GWAS
260 enrichment on OCRs from mouse non-brain tissues (kidney, liver, and lung) **(Supplemental**
261 **Figure 4D)**. As expected, we did not find an enrichment for genetic variants associated with a

262 brain-related trait. We did find that human orthologs of lung OCRs enrich for BMD, which concords
263 with the enrichment of human lung OCRs.
264
265 **Machine learning models of mouse cell type-specific CRE activity refine human NeuN+**
266 **OCRs for GWAS enrichment**
267 Assays such as ATAC-seq capture the open chromatin genomic regions that can be
268 occupied by DNA-binding regulatory proteins. DNA-binding proteins and their cofactors regulate
269 gene expression, often in response to external stimuli; therefore, the DNA sequences underlying
270 OCRs reveal the complex combinations of DNA-binding protein motif, i.e. the "*cis*-regulatory
271 grammar," that is being read by regulatory proteins to define the cell's epigenetic state. We
272 hypothesized that conservation of OCRs between mouse and humans must also rely on a conserved
273 regulatory grammar to maintain the similar cell identity between human and mouse (Chen et al.,
274 2018). The concordant pattern of enrichment for addiction associated genetic variants in human
275 and mouse-human orthologous OCRs suggested that risk variants may affect the regulatory
276 grammar of cell types conserved in mouse. We therefore trained a collection of convolutional
277 neural network (CNN) machine learning (ML) models to learn the regulatory grammar for cortical
278 excitatory (EXC) neurons, striatal D1 MSNs, and striatal D2 MSNs (Zhou and Troyanskaya, 2015;
279 Kelley et al., 2016, 2018; Chen et al., 2018). For each set of reproducible OCRs from mouse
280 INTACT and cSNAIL group, we trained 5-fold cross-validated models to predict the reproducible
281 peaks from ten times the number of GC-matched negative sequences **(Methods)**. Our models made
282 confident predictions on held-out test sequences as reported by accuracy, F1-score, and areas under
283 the receiver-operator characteristic and precision-recall curves **(Supplemental Figure 5A)**.

284 We reasoned that NeuN+ OCR signal that is comprised of several neuronal subtypes that
285 can be parsed into component cell types by ML models trained to predict OCR activity in those
286 cell types. To discern whether NeuN+ OCR enrichments in addiction-associated genetic variants
287 come from the same cell types that we see in **Figure 3**, we used the ML models to predict whether
288 cortical or striatal NeuN+ OCRs have activity in cortical EXC or striatal D1 and D2 cells,
289 respectively **(Figure 3A)**. We did not conduct these analyses for PV neurons because they
290 comprise a much lower percentage of cortical and striatal neurons than the other neuron types. We
291 ran LDSC regression (Finucane et al., 2018) GWAS enrichments on the sets of NeuN+ OCRs
292 predicted to be specific to cortical EXC, striatal D1, and striatal D2 neurons. Genetic variants
293 associated with `SmokingInitiation`, which initially enriched in OCRs of various NeuN+
294 frontal cortical areas **(Figure 1B)**, enriched in NeuN+ OCRs predicted to be active in EXC neurons
295 **(Figure 3B)**. Genetic variants associated with `Cannabis`, which enriched in NeuN+ cortical
296 OCRs **(Figure 1B)**, also enriched in NeuN+ OCRs predicted to be active in EXC neurons. The
297 enrichments of excitatory cortical cell type-specific OCRs for `SmokingInitiation` and
298 `Cannabis` associated genetic variants agree with the results from the Fullard *et al*. and Lake *et*
299 *al*. human datasets **(Figure 1B, C)**. Genetic variants associated with `SmokingCessation` and
300 `DrinksPerWeek`, which enriched in PUT and NAc NeuN+ OCRs **(Figure 1B)**, shared
301 enrichment in OCRs predicted active in both D1 and D2 MSNs of both PUT and NAc. The
302 framework that we outline in **Figure 3A** refines addiction genetic risk signal to neuronal subtypes,
303 maintains the brain region context of the source NeuN+ OCR, and can be applied to CREs from
304 any tissue, cell type combination for which bulk tissue open chromatin exists in human and open
305 chromatin from that cell type exists in another vertebrate(Chen et al., 2018; Minnoye et al., 2020).
306

307 **Machine learning models predict allele-specific activity of addiction-associated GWAS SNPs**
308 **in neuronal subtypes**

309 Lastly, we applied our ML models to screen addiction-associated genetic variants for
310 predicted functional activity in EXC, D1, and D2 neuronal subtypes. These ML-based approaches
311 have been demonstrated to fine-map dense risk loci and select candidate causal genetic variants
312 (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015; Kelley et al., 2016, 2018), yet none have
313 been applied in the context of addiction-associated genetic risk or in the cell types that we have
314 assayed. We collected 14,790 SNPs across the seven addiction-associated GWAS and filtered
315 down to 170 SNPs that are in NeuN+ OCRs and predicted to be causal with statistical fine-mapping
316 **(Methods)**. We demonstrated that our cortical and striatal ML models produce scores near 0 (no
317 predicted open chromatin activity in this cell type) for SNPs not in any OCRs, low scores for SNPs
318 in NeuN- OCRs, and significantly larger scores (larger predicted open chromatin in this cell type)
319 for SNPs in NeuN+ OCRs **(**$P_{Bonferroni} < 0.05$, **Figure 4A)**. Using these models and the 170 filtered
320 SNPs, we prioritized 26 unique SNPs spanning 16 loci with predicted functional consequence on
321 gene regulation **(Supplemental Figure 5B, Methods)**. One such SNP, rs7604640, lies in NeuN+
322 open chromatin specific to striatum ~46kb upstream of the *SIX3* locus on chromosome 2.
323 rs7604640 overlaps human orthologs of mouse OCRs in only D1 and D2 neurons and had
324 predicted open chromatin activity in both D1 and D2 neurons but not in EXC neurons **(Figure**
325 **4B)**. rs7604640 is one of many off-lead SNPs identified in the `SmokingInitiation` GWAS
326 ($P_{GWAS} = 3.04 \times 10^{-12}$) and is in LD with the lead SNP rs163522 ($R^2 = 0.856$). Furthermore, this
327 SNP is a known *cis*-eQTL for the antisense *SIX3-AS1* gene in striatal regions from the Genotype-
328 Tissue Expression (GTEX) project (GTEx Consortium, 2013, 2015; Melé et al., 2015; GTEx
329 Consortium et al., 2017). Anti-sense gene expression is one mechanism of regulating their sense

330 gene(Pelechano and Steinmetz, 2013; Barman et al., 2019), and deletion of the gene *SIX3* has been
331 shown to inhibit development of D2 medium spiny neurons (Xu et al., 2018). Altogether, this
332 evidence formulates the hypothesis that common genetic variant rs7604640 has D1 and D2 MSN-
333 specific open chromatin activity in a mouse-human conserved putative CRE regulating the MSN
334 regulator *SIX3*.

335 We found a number of other SNP candidates that may be putative causal SNPs with cell
336 type-specific activity in addiction-associated traits **(Supplemental Figure 6)**. For example, SNPs
337 rs11191352 ($P_{SmokingInitiation}$=2.12 x 10^{-7}) and rs9844736 ($P_{RiskyBehavior}$= 3.04 x 10^{-7},
338 $P_{SmokingInitiation}$=3.58 x 10^{-7}) are both *cis*-eQTL for nearby genes in GTEx striatal regions
339 **(Supplemental Figure 6A,B)**. Curiously, SNPs such as rs7604640, rs10742814, and rs11191352
340 had predicted cell type-specific activity in only MSNs **(Figure 4B, Supplemental Figure 6A,B)**,
341 while the SNP rs9844736 had predicted activity in mouse-human orthologous CREs from all three
342 cell types **(Supplemental Figure 6D)**. SNPs rs6870603 ($P_{SmokingInitiation}$=1.04 x 10^{-8}) and
343 rs7712167 ($P_{SmokingInitiation}$=8.79 x 10^{-9}), which are 317bp apart, were predicted to alter only EXC
344 open chromatin activity and lie in a strong cortical-specific NeuN+ OCR. The complete summary
345 of prediction scores and annotations of GTEX eQTL membership for each SNP can be found in
346 **Supplemental Table 1**.

347 **DISCUSSION**

348 In this study, we demonstrate the first analyses integrating cell type OCRs across human
349 and mouse brain epigenomics using ML models to select candidate addiction-associated SNPs
350 acting at putative cell type-specific CREs. We trained ML models to predict cell type-specific
351 activity of OCRs and used the models to predict whether addiction-associated genetic variants in
352 risk loci impact putative CRE function. Our findings link the genetic heritability of addiction-
353 associated behaviors to the OCR profiles of neuronal subtypes and brain regions and present
354 specific hypotheses describing how genetic variants may impact gene regulation in addiction-
355 associated behaviors. These analyses in conjunction suggest that genetic variation-associated
356 nicotine, alcohol, and cannabis use behaviors may impact putative CREs in different combinations
357 of excitatory (EXC), D1, and D2 neuronal subtypes. These findings provide a foundation for future
358 investigations into the cell type-specific genetic mechanisms underlying addiction-related traits.
359 More broadly, our cross-species integrative computational framework leverages high-resolution
360 cell-type targeted epigenomics in model organisms to interpret the genetic risk variants of complex
361 traits in humans.

362 We initially found that addiction-associated genetic variants were enriched in human
363 NeuN+ OCRs of the prefrontal cortex and striatum, known areas involved in addiction and reward
364 circuitry(Volkow et al., 2013; Koob and Volkow, 2016) **(Figure 5A)**. Genetic variants associated
365 with `SmokingInitiation` and `Cannabis`, initiating behaviors of substance use, were
366 enriched in NeuN+ OCRs of prefrontal areas including DLPFC, VLPFC, and OFC **(Figure 1B)**.
367 These OCRs were predicted to be active in cortical excitatory neurons of these brain regions
368 **(Figure 3B)**. Addiction-associated genetic variants that enrich in OCRs of cortical excitatory
369 neurons in these areas may reduce corticostriatal activation from prefrontal cortex to inhibit

370 behaviors predisposing the initiation of substance use(Koob and Volkow, 2010, 2016; Volkow et
371 al., 2013; Volkow and Morales, 2015). These genetic variants may contribute to reduced prefrontal
372 self-control reward, leading to behaviors observed in addiction such as impulsivity, reduced
373 satiety, and enhanced motivation to procure drugs(Volkow et al., 2013; Volkow and Morales,
374 2015). In addition, we found enrichment of striatal NeuN+ OCRs for genetic variants associated
375 with `SmokingCessation` and `DrinksPerWeek` **(Figure 1B)**. In **Figure 3B**, we showed that
376 these genetic variants are predicted to affect open chromatin in both D1 and D2 MSNs, which are
377 coordinators of mesocorticostriatal dopamine systems(Koob and Volkow, 2010, 2016; Volkow et
378 al., 2013). Genetic variants affecting open chromatin in these MSN subtypes may predispose
379 individuals to increased alcohol use (`DrinksPerWeek`) or decreased nicotine use
380 (`SmokingCessation`), perhaps driving the neuroplastic changes in D1 and D2 MSNs
381 observed in human and rodent drug dependence studies(Volkow et al., 1996, 1997, 2003; Wang et
382 al., 1997; Fehr et al., 2008; Cheng et al., 2017; Wilar et al., 2019). While drug addiction has been
383 attributed to various areas of the reward circuit, our investigations into heritable genetic risk for
384 addiction-associated traits unravel how regulatory DNA sequence variation in OCRs of projection
385 neurons in implicated areas link genetic risk to neural circuits to behavior.

386 Since cell types in the occipital cortex are not clearly defined for their role in the reward
387 circuit, we conducted ATAC-seq of projection and interneuron subtypes in mouse brain, mapped
388 OCRs to human orthologs, and analyzed GWAS enrichment of addiction-associated traits. By
389 leveraging ortholog mapping tools, we retained high-quality cell type-specific measurements
390 within relevant brain regions of the reward circuit, enabling analysis of cell populations from brain
391 regions where we lack primary human open chromatin profiles. Across these brain regions, we
392 found remarkably concordant enrichments of cell type OCRs between mouse and human profiles

393 as well as shared enrichments between traits **(Figure 5B)**. Genetic variants associated with both
394 `SmokingInitiation` and `Cannabis` enriched in mouse-human orthologous OCRs of
395 cortical EXC and PV neurons **(Figure 3B)**, concordant with enrichments in human cortical NeuN+
396 OCRs **(Figure 1B)**, which were predicted to include EXC neurons **(Figure 4B)**. Genetic variants
397 from these two traits showed replicable enrichment in human EXC and IN neuron OCRs via sc-
398 THS-seq of occipital cortex **(Figure 1C**), providing strong evidence that genetic variation in
399 cortical excitatory and inhibitory neuron OCRs confers susceptibility to nicotine and cannabis use
400 behaviors. Within striatal regions, D1 and D2 MSN mouse-human orthologous OCRs enriched for
401 genetic variants of all measured addiction-associated traits **(Figure 2B)**, with strongest
402 concordance in human OCRs for genetic variants associated with `SmokingCessation` and
403 `DrinksPerWeek` **(Figure 3B, Figure 5B)**. The enrichments in conserved OCRs of MSN
404 subtypes in the dorsal striatum and nucleus accumbens unsurprisingly emphasize known roles of
405 MSNs of both areas to drive and maintain addiction behaviors(Ferguson et al., 2011; Ji et al.,
406 2017).

407 In an orthogonal approach to mapping mouse-human orthologous OCRs, we leveraged
408 machine learning models to learn the regulatory grammar of neuronal subtypes characterized in
409 mouse and refine human NeuN+ OCRs to the major neuronal subtypes of cortex and striatum for
410 GWAS enrichment. Refinement of NeuN+ OCRs revealed that addiction-associated traits enriched
411 for two clusters of cell types and brain regions. The first group, which displays concordant cortical
412 excitatory enrichments between human and mouse, consists of `SmokingInitiation` and
413 `Cannabis` **(Figure 3B)**, and the second group, which displays concordant D1 and D2 MSN
414 enrichments, consists of `SmokingCessation` and `DrinksPerWeek`. A draw-back of
415 assigning human NeuN+ OCR membership to individual cell types lies in the considerably low

416 representation of interneurons in both cortical and striatal neuron populations - as low as 12% in
417 neocortex(Beaulieu, 1993; Lefort et al., 2009) and ~5% in striatum(Tepper and Koós, 2017;
418 Krienen et al., 2019). NeuN+ open chromatin profiles alone do not always capture OCRs unique
419 to rare interneurons, some of which had OCRs identified by human single-cell assays and mouse-
420 human orthologs enriched for addiction GWAS variants **(Figure 1C, Figure 2B)**. As a result, we
421 did not train ML models for PV, SST, or VIP interneurons. However, the striking enrichments of
422 OCRs from certain interneuron populations for addiction GWAS variants begin to demonstrate
423 these populations' roles in the addiction neural circuits(Bracci et al., 2002; Lansink et al., 2010;
424 Wiltschko et al., 2010; Ribeiro et al., 2018; Jiang et al., 2019; Lee et al., 2020; Schall et al., 2020).
425 The overall concordance of enrichments across human and mouse-human orthologous
426 OCRs suggests a conserved regulatory grammar between mouse and human CREs.
427 Correspondence in the neural circuitry has been well-appreciated between human studies and
428 rodent models of addiction(Berke and Hyman, 2000; Koob and Volkow, 2016; Farrell et al., 2018),
429 and our study further demonstrates that mouse-human conserved OCRs may explain considerable
430 heritability of addiction-associated traits. This makes animal models suitable not only for studying
431 the neural circuits of addiction but also cell-type-specific gene-regulatory mechanisms of
432 addiction.
433 We used several selection criteria along with ML models to predict the functional impact
434 of genetic variants that may be agents in addiction-associated traits **(Figure 5, Supplemental**
435 **Figure 5, Supplemental Table 1)**. The fine-mapping pipeline described effectively narrows down
436 a set of 14,790 SNPs to a putatively functional set of 26 candidates that can be experimentally
437 tested to determine which brain regions and neuronal subtypes they would have function in. The
438 candidate functional SNPs that our models prioritize demonstrate how a candidate SNP within a

439 locus, such as rs7604640 **(Figure 4B)**, might act in different combinations of neuronal subtypes
440 and brain regions. This pleiotropy adds complexity to discerning which neuronal subtypes have
441 altered gene regulation in addiction-associated traits. Our approach often reported only one or two
442 candidates per loci, reflecting the idea that many SNPs in the same loci are significantly associated
443 with addiction due to LD with the causal SNP and may not influence addiction-associated genetic
444 predisposition. The candidate SNPs that overlap mouse-human orthologs from the same predicted
445 cell type raise the idea that altering the conserved regulatory DNA sequence may be a mechanism
446 of cell type-specific gene regulatory tuning in a population or even across species(Gjoneska et al.,
447 2015).

448 Our study depends solely on assays of open chromatin as a proxy for putative CREs.
449 Epigenetic assays for chromatin conformation, histone modifications, and methylation would
450 further inform how putative CREs regulate nearby gene expression. Furthermore, our predictions
451 of SNP impact on putative CREs and genes remain to be experimentally validated. While eQTL
452 studies do not control for inflated associations due to LD and report gene expression differences
453 from bulk tissue, we do note that our approach prioritizes several SNPs known to be *cis*-eQTLs in
454 relevant brain regions, which indirectly affirms our framework's ability to select SNPs with
455 functional impacts on gene regulation. In order to validate our predictions, it will be necessary to
456 further investigate candidate genetic variants such as rs7604640 **(Figure 4B)** in future studies
457 using massively parallel reporter assays(Tewhey et al., 2016) or self-transcribing active regulatory
458 region sequencing(Vockley et al., 2015; Kalita et al., 2018) that can measure regulatory activity
459 differences between risk and non-risk alleles. The candidate SNPs we identified provide possible
460 mechanisms linking differences in genetic make-up with the genes, cell types, and brain regions
461 that could influence addiction and substance use behaviors **(Figure 4)**.

462 **MATERIALS & METHODS**

463 **ATAC-seq data processing pipeline:**

464 We processed raw FASTQ files of ATAC-seq experiments with the official ENCODE ATAC-seq
465 pipeline (Landt et al., 2012) accessed by https://github.com/ENCODE-DCC/atac-
466 seq-pipeline. We ran this pipeline using the mm10 genome assembly for mouse and the hg38
467 genome for human with the following settings: smooth_win = 150, multimapping = 0,
468 idr_thresh = 0.1, cap_num_peak = 300,000,
469 keep_irregular_chr_in_bfilt_peak = true. We grouped biological replicates when
470 processing data to obtain individual de-duplicated, filtered bam files and reproducible (IDR) peaks
471 for each condition. Unless otherwise stated, we used the optimal reproducible set of peaks for
472 downstream analyses. We removed samples that had low periodicity indicated by ENCODE
473 quality control metrics and reprocessed the remaining replicates with the pipeline.

474 **Publicly available datasets**

475 Fullard *et al.* NeuN-sorted ATAC-seq of human postmortem brain (Fullard et al., 2018): We
476 identified OCRs overlapping addiction-related variants through analysis of human postmortem
477 brain ATAC-seq in which cells were sorted into NeuN-positive and NeuN-negative groups via
478 fluorescence activated nuclei sorting (FANS); the brain regions we used were dorsolateral
479 prefrontal cortex (DLPFC), orbitofrontal cortex (OFC), ventrolateral prefrontal cortex (VLPFC),
480 anterior cingulate cortex (ACC), superior temporal gyrus (STC), inferior temporal gyrus (ITC),
481 primary motor cortex (PMC), insula (INS), primary visual cortex (PVC), amygdala (AMY),
482 hippocampus (HIP), mediodorsal thalamus (MDT), nucleus accumbens (NAc), and putamen
483 (PUT). We downloaded data from the Sequence Read Archive (SRA) through Gene Expression
484 Omnibus (GEO) accession #GSE96949. We separated samples by cell type and reprocessed them

485 with the ENCODE pipeline as detailed above, aligning reads to hg38. We used the "optimal
486 reproducible peaks" for each cell type and brain region as foregrounds in GWAS LDSC
487 enrichment with the Honeybadger2 OCR set as the background set (see LDSC Regression GWAS
488 Enrichment Backgrounds).
489 Lake et al. human occipital cortex scTHS-seq (Lake et al., 2018): We downloaded BED-formatted
490 cell type-specific differential OCRs from occipital cortex scTHS-seq of excitatory neurons (EXC),
491 inhibitory neurons (IN), astrocytes (AST), endothelial cells (END), oligodendrocyte precursor
492 cells (OPC), oligodendrocytes (OLI), and microglia (MIC) from the GEO subseries #GSE97887.
493 We used the hg38 OCR coordinates as foregrounds in LDSC regression GWAS enrichment with
494 the Honeybadger2 OCR set as the background set (LDSC regression GWAS Enrichment
495 Backgrounds).
496 Mo *et al.* mouse INTACT-sorted nuclei ATAC-seq (Mo et al., 2015): We downloaded FASTQ
497 files of *R26-CAG-LSL-Sun1-sfGFP-Myc* transgenic mouse lines for cell type-specific ATAC-seq
498 performed using the INTACT method from the accession #GSE63137. *Mo et al.* isolated INTACT-
499 enriched nuclei from three cell types: excitatory neurons (EXC, *Camk2a-cre*), vasoactive intestinal
500 peptide neurons (VIP, *Vip-cre*), and parvalbumin neurons (PV, *Pvalb-cre*). We reprocessed the
501 data with the Kundaje Lab open chromatin pipeline using the mm10 genome
502 (https://github.com/kundajelab/atac_dnase_pipelines). We mapped
503 reproducible mouse ATAC-seq peaks for each cell type to hg38 using halLiftover with the 12-
504 mammals Cactus alignment (Paten et al., 2011; Hickey et al., 2013) followed by HALPER (Zhang
505 et al., 2020) (Mapping mouse OCR orthologs) to produce a foreground set of orthologous human
506 sequences for LDSC regression GWAS enrichment (Finucane et al., 2018). We mapped the
507 ENCODE mm10 DNaseI-hypersensitive peak set(Yue et al., 2014) to hg38 (Mapping mouse OCR

508 orthologs) and used successfully mapped hg38 orthologs of mm10 OCRs a background set for
509 mouse foreground enrichments. Furthermore, we used this dataset to evaluate differential
510 accessibility in cSNAIL-INTACT PV and PV-negative ATAC-seq samples and develop machine
511 learning models of cell type-specific open chromatin (see Methods below).
512 Human negative control foregrounds (ENCODE Project Consortium, 2012; Thurman et al., 2012;
513 Davis et al., 2018; Cannon et al., 2019): We downloaded raw ATAC-seq profiles of human adult
514 female and male stomach ATAC-seq generated by Snyder *et al*. (ENCSR337UIU,
515 ENCSR851SBY, respectively), female human embryonic liver DNase-seq generated by
516 Stamatoyannopoulos *et al*. (ENCSR562FNN), and human embryonic lung DNase-seq generated
517 by Stamatoyannopoulos *et al*. (ENCSR582IPV) from
518 `https://www.encodeproject.org/`. We processed these files using the ENCODE
519 pipeline as detailed above to obtain optimal reproducible `hg38` peaks. We also downloaded BED
520 files of human adipocyte and preadipocyte ATAC-seq profiles generated by Cannon *et al*. from
521 GEO accession number #GSE110734. We mapped these BED coordinates from `hg19` to `hg38`
522 using liftOver to define negative control foregrounds for human LDSC regression GWAS
523 enrichment. We merged the human negative control foregrounds and Fullard *et al.* foregrounds
524 with the Honeybadger2 OCR set to define the background for human negative control foreground
525 enrichments.
526 Human-orthologous negative control foregrounds (Liu et al., 2019a): We also downloaded raw
527 ATAC-seq data profiled in female mouse kidney, female mouse liver, and male mouse lung
528 generated by Liu *et al.* from SRA accession #SRP167062 to define human-orthologous negative
529 control foregrounds. We processed these files using the ENCODE pipeline as detailed above to
530 get optimal reproducible peaks. We mapped optimal reproducible peaks from `mm10` to `hg38` using

531 halLiftover with the 12-mammals Cactus alignment followed by HALPER (Mapping mouse OCR
532 orthologs) to define negative control foregrounds for human-orthologous LDSC GWAS
533 enrichments. We merged all human orthologous foregrounds with the human orthologs of the
534 ENCODE mm10 DNaseI-hypersensitive peak set to define a background for human-orthologous
535 LDSC GWAS enrichments.

536 **Mapping mouse open chromatin region (OCR) orthologs**

537 We employed halLiftover (Hickey et al., 2013) with the 12-mammals Cactus alignment (Paten et
538 al., 2011) followed by HALPER
539 (https://github.com/pfenninglab/halLiftover-postprocessing)(Zhang et
540 al., 2020) to map mm10 mouse reproducible OCRs to hg38 human orthologs in order to perform
541 LDSC regression GWAS enrichment. The Cactus multiple sequence alignment file (Paten et al.,
542 2011) has 12 genomes, including mm10 and hg38, aligned in a reference-free manner, allowing
543 us to leverage multi-species alignments to confidently identify orthologous regions across species.
544 halLiftover uses a Cactus-format multiple species alignment to map BED coordinates of a query
545 species to orthologous coordinates of a target species, and HALPER constructs contiguous
546 orthologs from the outputs of halLiftover (Zhang et al., 2020). We ran the orthologFind.py
547 function from HALPER on the outputs of halLiftover using the following parameters: -
548 max_frac 5.0 -min_frac 0.05 -protect_dist 5 -narrowPeak -
549 mult_keepone. In general, 70% of mouse brain ATAC-seq reproducible peaks were able to be
550 mapped to confident human orthologs. To map the ENCODE mm10 mouse DHS background,
551 which does not contain summit information, to hg38 we used the mouse coordinates of position
552 with the most species aligned in a region to define the summit. Only for the mm10 mouse DHS
553 background set, for which a significant proportion of regions could not be confidently mapped to

554 hg38, we flanked the original assembly coordinates by 300 bp to increase OCR mapping from
555 54% to 64%.

556 **LDSC Regression GWAS Enrichment Backgrounds:**

557 We found that LDSC regression GWAS enrichment analysis is sensitive to the selected
558 background set of matched regions. To construct appropriate background sets for each GWAS
559 enrichment, we used the ENCODE and RoadMap Honeybadger2(Roadmap Epigenomics
560 Consortium et al., 2015) and Mouse DHS peak sets for the respective human and mouse-based
561 open chromatin GWAS enrichment. The Honeybadger2 set, downloaded from
562 https://personal.broadinstitute.org/meuleman/reg2map/, consists of
563 DNaseI-hypersensitive OCRs across 53 epigenomes consisting of promoter, enhancer, and dyadic
564 regions. Honeybadger2 is an appropriate epigenetic reference for enrichment of cell type-specific
565 open chromatin from various foregrounds such as the Fullard *et al.* and Lake *et al*. Honeybadger2
566 regions allow the LDSC algorithm to properly account for the heritability from OCRs of a
567 particular cell type or regions rather than because they tend to be more conserved, are enriched for
568 ubiquitously active transcription factor motifs, or other factors distinguishing open chromatin from
569 heterochromatin. The human orthologs of the ENCODE Mouse DHS peak set, downloaded
570 through the ENCODE ATAC-seq pipeline at
571 http://mitra.stanford.edu/kundaje/genome_data/mm10/ataqc/mm10_un
572 iv_dhs_ucsc.bed.gz, is a set of peaks combined from mouse DNaseI-hypersensitivity
573 OCRs from ENCODE and provides a background for human orthologs of mouse OCRs. The mm10
574 mouse DHS regions were mapped to hg38 as described in **Mapping mouse OCR orthologs**. For
575 each respective foreground-background pairing, the foreground regions were merged with the
576 background reference to ensure the background always contained the foreground set. The mouse

577 background was used to calculate the significance of the relationship between mouse OCRs and
578 GWAS for addiction-associated traits to control for a possible association between the degree to
579 which a region is conserved and its likelihood in influencing the predisposition to an addiction-
580 associated trait.

581 **GWAS enrichment with partitioned LD score regression analysis**

582 We computed the partitioned heritability of CREs for GWAS variants using the LDSC regression
583 pipeline for cell type-specific enrichment as outlined in
584 `https://github.com/bulik/ldsc/wiki/Cell-type-specific-`
585 `analyses`(Bulik-Sullivan et al., 2015b). We downloaded the GWAS summary statistics files and
586 processed them with the LDSC `munge_sumstats` function to filter rare or poorly imputed SNPs
587 with default parameters. We munged the summary statistics files for HapMap3 SNPs excluding
588 the MHC regions downloaded at
589 `http://ldsc.broadinstitute.org/static/media/w_hm3.noMHC.snplist.`
590 `zip`. We inspected GWAS file to ensure the effect allele, non-effect allele, sample size, p-value,
591 and signed summary statistic for each SNP in each GWAS were included and appropriate for
592 LDSC. The addiction-associated GWAS measure genetic predisposition for age of smoking
593 initiation (`AgeofInitiation`)(Liu et al., 2019b), heaviness of smoking
594 (`CigarettesPerDay`)(Liu et al., 2019b), having ever regularly smoked
595 (`SmokingInitiation`)(Liu et al., 2019b), current versus former smokers
596 (`SmokingCessation`)(Liu et al., 2019b) , alcoholic drinks per week (`DrinksPerWeek`)(Liu
597 et al., 2019b), cannabis consumption (`Cannabis`)(Pasman et al., 2018), and risk tolerance
598 (`RiskyBehavior`)(Karlsson Linnér et al., 2019). GWAS traits related to addiction include
599 multisite chronic pain (`ChronicPain`)(Johnston et al., 2019) and number of coffee cups drank

per data (CoffeePerDay)(Coffee and Caffeine Genetics Consortium et al., 2015). Other addiction-related traits come from underpowered GWAS including opioid dependence (OpioidDep)(Cheng et al., 2018) , cocaine dependence (CocaineDep)(Cabana-Domínguez et al., 2019), and diagnosis of obsessive-compulsive disorder (OCD)(International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS), 2018). GWAS from strong brain-related traits used are schizophrenia risk (Schizophrenia)(Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), highest level of educational attainment (EduAttain)(Lee et al., 2018), and sleep duration (SleepDuration)(Dashti et al., 2019). The non-brain related traits measure genetic liability for lean body mass (LBM)(Zillikens et al., 2017), bone mineral density (BMD)(Kemp et al., 2017), and coronary artery disease (CAD)(Howson et al., 2017).

We estimated LD scores for each foreground set and corresponding background set with the LDSC regression pipeline make_annot and ldsc functions using hg38 1000 Genomes European Phase 3 European super-population (1000G EUR) cohort plink files downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/GRCh38/. An example of an ATAC-seq optimal set of reproducible peaks mapped to hg38 in narrowPeak format is annotated with 1000G EUR LD scores using the following call:

```
python make_annot.py \
--bed-file optimal_peak.narrowPeak.gz \
--bimfile 1000G.EUR.hg38.${chr}.bim \
--annot-file foreground.${chr}.annot
```

We downloaded the baseline v1.2 files for cell type-specific enrichment in hg38 coordinates from the same link above as well as the corresponding weights

623 `weights.hm3_noMHC' file excluding the MHC region from
624 https://data.broadinstitute.org/alkesgroup/LDSCORE/. HapMap SNPs and
625 corresponding weights file used in the LDSC analyses only refer to the SNP rsIDs, rather than
626 genomic coordinates, so only the baseline and LD statistics used to annotate the foreground and
627 background files must be in hg38 coordinates. In accordance with the LDSC regression script
628 input format, we created an `enrichment.ldcts' file listing the annotated
629 foreground/background pair for each foreground set. We estimated the partitioned heritability
630 using the ldsc function, which integrates the foreground and background LD score estimates,
631 munged GWAS SNP data, baseline variant data, and variants weights. The final function call to
632 GWAS enrichment is as follows:

```
633 python ldsc.py --h2-cts $Munged_GWAS \
634       --ref-ld-chr baseline_v1.2/baseline. \
635       --w-ld-chr weights.hm3_noMHC. \
636       --ref-ld-chr-cts enrichment.ldcts \
637       --out $Output_Label
```

638 The pipeline produced LD score regression coefficient, coefficient error, and coefficient p-
639 value estimates. We adjusted for multiple testing using the false discovery rate on p-values of the
640 LD score regression coefficients (alpha = 0.05) on all 18 GWAS traits intersected on within
641 the same foreground/background set. A significant FDR-value indicates enrichment of the
642 foreground genomic regions for GWAS SNPs relative to the background. Lastly, we computed
643 genetic correlations in **Supplemental Figure 1A** between GWAS of addiction-associated traits
644 using the pre-munged summary statistics as described by Bulik-Sullivan *et al.*(Bulik-Sullivan et
645 al., 2015a)

646 **Bulk tissue ATAC-seq**

647 To augment and compare to mouse cell type-specific ATAC-seq datasets generated in this study,
648 we also performed bulk tissue ATAC-seq from cortex (CTX) and dorsal striatum/nucleus
649 accumbens (CPU) of 7- and 12-week-old C57Bl/6J mice (N = 2 each age) as described in
650 Buenrostro *et al.*, 2015(Buenrostro et al., 2015) with the following minor differences in buffers
651 and reagents. We euthanized mice with isoflurane, rapidly decapitated to extract the brain, and
652 sectioned it in ice-cold oxygenated aCSF (119mM NaCl, 2.5 mM KCl, 1mM
653 NaH_2PO_4(monobasic), 26.2mM $NaHCO_3$, 11mM glucose) at 200-micron sections on a vibratome
654 (Leica VT1200). We further micro-dissected sections for cortex and dorsal striatum on a stereo
655 microscope and transferred dissected regions into chilled lysis buffer (Buenrostro et al., 2015). We
656 dounce homogenized the dissected brains in 5mL of lysis buffer with the loose pestle (pestle A) in
657 a 15mL glass dounce homogenizer (Pyrex #7722-15). We washed nuclei lysate off the pestle with
658 5mL of lysis buffer and filtered the nuclei through a 70-micron cell strainer into a 50mL conical
659 tube. We washed the dounce homogenizer again with 10mL of BL buffer and transferred the lysate
660 through the 70-micron filter (Foxx 1170C02). We pelleted the 20 mL of nuclei lysate at 2,000 x *g*
661 for 10 minutes in a refrigerated centrifuge at 4°C. We discarded the supernatant and resuspended
662 the nuclei in 100-300 microliters of water to approximate a concentration of 1-2 million nuclei/
663 mL. We filtered the nuclei suspension through a 40-micron cell strainer. We stained a sample of
664 nuclei with DAPI (Invitrogen #D1206) and counted the sample to measure 50k nuclei per ATAC-
665 seq transposition reaction. The remaining steps follow the Buenrostro *et al.*, 2015(Buenrostro et
666 al., 2015) protocol for tagmentation and library amplification. We shallowly sequenced barcoded
667 ATAC-seq libraries at 1-5 million reads per sample on an Illumina MiSeq and processed individual
668 samples through the ENCODE pipeline for initial quality control. We used these QC measures

669 (clear periodicity, library complexity, and minimal bottlenecking) to filter out low-quality samples
670 and re-pooled a balanced library for paired-end deep sequencing on an Illumina NextSeq to target
671 30 million uniquely mapped fragments per sample after mitochondrial DNA and PCR duplicate
672 removal. These raw sequencing files entered processing through the ENCODE ATAC-seq pipeline
673 as above by merging technical replicates and grouping biological replicates by brain region for
674 each pipeline run.

675 **Cre-Specific Nuclear-Anchored Independent Labeling (cSNAIL) virus procedures**

676 The cSNAIL genome (pAAV-Ef1a-DIO-Sun1-Gfp-WPRE-pA) contains *loxP* sites to invert the
677 *Sun1-Gfp* fusion gene and integrate into the nuclear membrane of cells expressing the *Cre* gene,
678 allowing these cell populations to be profiled for various genomic assays (Lawler et al, 2020 in
679 press J. Neuro). We packaged the cSNAIL genome with AAV variant PHP.eB (pUCmini-iCAP-
680 PHP.eB) in AAVpro(R) 293T cells (Takara, cat #632273). Viviana Gradinaru provided us with
681 the pUCmini-iCAP-PHP.eB (http://n2t.net/addgene:103005; RRID: Addgene 103005)(Chan et al.,
682 2017). We precipitated viral particles with polyethylene glycol, isolated with ultracentrifugation
683 on an iodixanol density gradient, and purified in PBS with centrifugation washes and 0.2µM
684 syringe filtration. We injected each mouse with 4.0 x 10^{11}vg into the retro-orbital cavity under
685 isoflurane anesthesia. We allowed the virus to incubate in the animal for 3-4 weeks to reach peak
686 expression. We closely monitored the health of the animals throughout the length of the virus
687 incubation and did not note any concerns.

688 **cSNAIL nuclei isolation**

689 On the day of the ATAC-seq experiments, we dissected brain regions from fresh tissue and
690 extracted nuclei in the same manner as described for bulk tissue experiments. Then, we sorted the
691 nuclei suspension into Sun1GFP+ (Cre+) and Sun1GFP- (Cre-) fractions using affinity purification

692 with Protein G Dynabeads (Thermo Fisher, cat. 10004D). A pre-clearing incubation with beads
693 and nuclei for 10-15 minutes removes effects from non-specific binding events. Next, we
694 incubated the remaining free nuclei with anti-GFP antibody (Invitrogen, #G10362) for 30 minutes
695 to bind Sun1GFP. Finally, we added new beads to the solution to conjugate with the antibody and
696 incubated the reaction for an additional 20 minutes. The pre-clear step and all incubations took
697 place in wash buffer (0.25M Sucrose, 25mM KCl, 5mM $MgCl_2$, 20mM Tricine with KOH to pH
698 7.8, and 0.4% IGEPAL) at 4°C with end-to-end rotation. After the binding process, we separated
699 bead-bound nuclei on a magnet, washed three times with wash buffer, and filtered through a 20µM
700 filter to ensure purity. We resuspended nuclei in nuclease-free water for input into the ATAC-seq
701 tagmentation reaction. We performed nuclei quantification and tagmentation in the same manner
702 described for bulk tissue ATAC-seq above. We list in the table below the number of animals, the
703 genotypes, and which regions collected for ATAC-seq experiments in this study. N=2 *Pvalb-cre*
704 samples from CPU/NAc region had received a sham surgery with saline injection into the external
705 globus pallidus 5 days before they were sacrificed (Lawler et al, 2020 in press J. Neuro.) N=2
706 *Drd1-cre* samples from both CPU and NAc regions had received headcap surgeries 3 weeks before
707 they were sacrificed. Both *Pvalb-cre* and *Drd1-cre* were overall healthy at time of sacrifice.

Genotype	**Replicates**	**Sex** (**F**emale /**M**ale)	**Region and Replicate per region**	**Cell type**
C57BL/6 WT	N=4	2 F, 2 M	CTX= 4, CPU/NAc = 4	bulk
Pvalb-cre	N=5	3 F (CTX) 1 F, 1 M (CPU/NAc)	CTX= 3, CPU/NAc = 2	PV
Sst-cre	N=2	1 F, 1 M	CTX= 2, CPU/NAc = 2	SST
Drd1-cre	N=2	2 F	CPU=2, NAc=2	D1
Adora2a-cre	N=2	2 F	CPU=2, NAc=2	D2

708

709 **cSNAIL Cell Type Specificity**

710 We created a consensus set of non-overlapping IDR peaks from the ATAC-seq pipeline for
711 cSNAIL ATAC-seq and Mo *et al.* INTACT samples (`Tissue`: Ctx, Cpu, and NAc ; `Celltype`:
712 EXC, PV, SST, VIP, D1, D2). We extended the peak set 200bp up- and down-stream, count
713 overlapping fragments with `Rsubread v2.0.1` using the de-duplicated BAM files from the
714 pipeline(Liao et al., 2014), and created with `DESeq2 v1.26.0` a variance-stabilized count
715 matrix aware of experimental `Group` (combination of `Tissue` and `Celltype`) with `~Group`
716 (Love et al., 2014). We plotted the principle component analysis in **Supplemental Figure 3A** for
717 the first two components with this variance-stabilized count matrix. We used `Deeptools`
718 `v3.5.0` to convert the same BAM files to normalized bigWig files and average over replicates
719 of the same `Group`(Ramírez et al., 2016), We plotted the tracks using `pyGenomeTracks v3.5`
720 around marker genes for each cell type (*Slc17a7*, *Drd1*, *Adora2a*, *Pvalb*, *Sst*, *Vip*)(Ramírez et al.,
721 2018) **Supplemental Figure 3B**. We computed the mean accessibility for each `Group` 2kb up-
722 and down-stream the transcription start sites (TSS) and correlated $\log_{10}$(TSS accessibility + 1) with
723 gene expression $\log_{10}$(meta gene counts + 1) of Drop-Seq annotated cell types from prefrontal
724 cortex and striatum(Saunders et al., 2018). We used the Saunders *et al.* tissue subcluster metagene
725 profiles (sum of gene expression in all cells) and summed subclusters to cluster-level metagene
726 profiles. Most tissue cluster metagene profiles corresponded to cSNAIL ATAC-seq celltype and
727 tissue profiles, with the exception of cSNAIL cortical PV+ samples were matched to Saunders *et*
728 *al.* cortical MGE+ interneuron clusters.
729

730 **<u>Convolutional Neural Network models for CRE cell type classification</u>**

731 We trained a set of convolutional neural network (CNN) models to learn the regulatory grammar
732 of a given cell type from the DNA sequences underlying the cell type's OCRs. The models take in

733 one-hot encoded 501bp genomic sequences, where positives are centered on the IDR peak summits
734 that are annotated to be in introns and distal intergenic regions and negatives are approximately
735 ten times the number of positives sequences that are G/C-matched and not overlapping IDR peaks,
736 to predict 1 for an OCR or 0 for negative sequence. We excluded promoters (defined as within
737 5,000bp from the TSS) and exons since distal sequences have been shown to confer more cell type-
738 specificity and be more predictive of expression levels of regulated genes (Roadmap Epigenomics
739 Consortium et al., 2015). We constructed the negative set by first building a sequence repository
740 `$BGDIR` according to
741 https://bitbucket.org/CBGR/biasaway_background_construction/src/
742 master/ from the mouse `mm10` genome using 501bp sequences. Then we used the
743 `biasaway`(Worsley Hunt et al., 2014; Khan et al., 2020) command-line interface-generated
744 negative sequences with the matching nucleotide distribution along a sliding window along the
745 501bp IDR peak sequence:

```
746 biasaway c --foreground $FGFASTA --nfold 10 --deviation 2.6 --step
747 50 --seed 1 -winlen 100 --bgdirectory $BGDIR
```

748 We employed a 5-fold cross validation chromosome hold-out scheme to train 5 models per set of
749 IDR peaks to ensure stable and consistently learned regulatory patterns. A model for training a
750 fold does not see sequences during training from the validation set for that fold, and no models see
751 the test set until final model performance evaluation. Sequences from these chromosomes were
752 used as the validation set for each fold:

753 fold1: {`chr6, chr13, chr21`}
754 fold2: {`chr7, chr14, chr18`}
755 fold3: {`chr11, chr17, chrX`}

fold4: {chr9, chr12}

fold5: {chr10, chr8}.

We used sequences from chromosomes {chr1, ch2, chr19} for the test set.

We trained the models with Keras v2.3.0-tf (https://keras.io/) implemented through Tensorflow v2.2.0 and used stochastic gradient descent (SGD) with Nesterov momentum to minimize the binary cross entropy loss and learn model parameters. All models used the same CNN architecture with five consecutive Conv2D layers (conv_width = 11, conv_height = 4, conv_filters = 200, stride = 1), one MaxPooling2D layer (max_pool_size = 26, max_pool_stride = 26), one Dense layer (dense_filters = 300), one Dropout layer (proportion_dropout = .1), and a final output Dense layer (activation = 'sigmoid'). All Conv2D and Dense used the 'relu' activation and L2 regularization (l2_reg = 1e-10), unless otherwise stated. We applied the One-Cycle-Policy (OCP) with linear cyclical learning rate and momentum between a base and max rates as described previously (Smith, 2018) to train each fold with batch size= 2000, epochs = 23, num_cycles = 2.35, base_learning_rate = 1e-2, max_learning_rate = 1e-1, base_momentum = .85, max_momentum = 0.99. These parameters robustly and efficiently trained models across folds to accurately predict positive OCRs of all cell types against an approximately 1:10 positive:negative class imbalance. We computed classifier performance metrics including weighted accuracy (using threshold = 0.5), weighted f1_score (using threshold = 0.5), area under receiver operating characteristic (auROC), and area under precision-recall curve (auPRC). We provide the scripts and .h5 files with Keras model architectures and weights on the GitHub page https://github.com/pfenninglab/addiction_gwas_enrichment.

779

780 **Machine learning cell type-specific prioritization of *Fullard et al.* NeuN+ ATAC-seq peaks**

781 We used CNN model scores to classify whether a peak from *Fullard et al*. NeuN+ open chromatin
782 data is active in a neuronal subtype {EXC, D1, D2}. We took NeuN+ IDR "optimal peaks"
783 from regions significantly enriched for addiction-associated traits {OFC, VLPFC, DLPFC,
784 ACC, STC, PUT, NAc, **Figure 1A**}, extracted 501bp DNA sequences of each centered on the
785 summit, and scored each peak with cell type-specific machine learning models trained with the
786 appropriate tissue context (e.g. score cortical NeuN+ peaks with a model trained with cortical EXC
787 cell type). We averaged scores across model folds from the same cell types and classified NeuN+
788 peaks with scores greater than 0.5 as active in that cell type. We defined these ML-prioritized
789 peaks as foregrounds for LDSC regression GWAS enrichment analyses as described above. We
790 created a consensus set of peaks merging all model-prioritized peaks and the Honeybadger2 set of
791 OCRs to be the matched background, and we performed GWAS enrichment and computed FDR
792 on all 18 GWAS traits (only enrichments for addiction-associated GWAS shown, **Figure 3**).

793

794 **Addiction-associated GWAS processing and cell type-specific candidate selection**

795 We collected the addiction-associated SNPs by submitting the summary statistics files for the
796 seven addiction-associated traits{AgeofInitiation (Liu et al., 2019b),
797 CigarettesPerDay (Liu et al., 2019b), SmokingInitiation (Liu et al., 2019b),
798 SmokingCessation (Liu et al., 2019b), DrinksPerWeek (Liu et al., 2019b), Cannabis
799 (Pasman et al., 2018), RiskyBehavior (Karlsson Linnér et al., 2019)} to the FUMA webserver
800 (Watanabe et al., 2017). FUMA computed LD R^2 based on the 1000 Genomes European (1000G
801 EUR) super-population reference (1000 Genomes Project Consortium et al., 2015) via PLINK

802 (Purcell et al., 2007), linked GWAS-significant lead SNPs to off-lead SNPs in LD with the lead,
803 and annotated functional consequences of genetic variants via ANNOVAR based on ENSEMBL
804 build 85 human gene annotations (Wang et al., 2010) **(Figure 1A)**. ANNOVAR functional gene
805 annotations for a SNP are as defined in the primary publication and online:
806 `https://annovar.openbioinformatics.org/en/latest/user-`
807 `guide/gene/`. The aggregate of 14,790 unique SNPs span 215 genomic loci. We limited
808 ourselves SNPs that overlapped Fullard *et al.* NeuN+ OCRs (Fullard et al., 2018) since nucleotide
809 variants in these peaks may disrupt epigenomic DNA sequences measured by ATAC-seq. We also
810 limited ourselves SNPs that are fine-mapped and predicted to be causal by CAUSALdb using the
811 European LD structure and an ensemble of statistical fine-mapping tools (FINEMAP,
812 CAVIARBF, PAINTOR) (Chen et al., 2015; Benner et al., 2016; Kichaev et al., 2017; Wang et
813 al., 2020). Combining these two filtering heuristics, being in NeuN+ OCRs and fine-mapped
814 putatively causal, narrowed us down to 170 SNPs over 54 loci to be further refined for cell type-
815 specific activity.

816 We further filtered the 170 SNPs to only those overlapping OCRs from cortical and striatal
817 brain regions are enriched for addiction-associated variants { `OFC, DLPFC, VLPFC, ACC,`
818 `STC, PUT, NAc`} (**Figure 1A**), and scored the filtered SNPs with ML models trained on mouse
819 cortical or striatal cell type-specific ATAC-seq. We inputted DNA sequences of 501 bp centered
820 at the SNP location for both the effect and non-effect allele into the ML models for predicting cell
821 type-specific OCR activity and, for each cell type, computed the average cell type score
822 (prediction) across models from different folds for each of the effect and non-effect alleles. To
823 predict a measure of functional impact of the effect allele, we computed a SNP delta score
824 ($score_{effect} - score_{non\text{-}effect}$). Most SNPs reported by GWAS are not expected to be the

825 causal variant for a trait, so the distribution of cell type-specific model scores on the full list of
826 14,790 SNPs can be used to define a null distribution. From the 170 SNPs, we predict whether a
827 SNP might have functional impact in cell type if either allele has a score > 0.5 and if the delta
828 score magnitude is > 0.05, the standard deviation of null delta scores. To accompany cell type-
829 specific activity predictions, we downloaded SNPs that are reported *cis* expression quantitative
830 trait loci (eQTL) in human cortex, frontal cortex (DLPFC), ACC, caudate, putamen, and NAc from
831 the GTEX Consortium from https://www.gtexportal.org/home/datasets(GTEx
832 Consortium, 2013, 2015). We identified genes for which at least one of the 170 SNPs is
833 an eQTL and plotted them as arcs in **Figure 4B** and **Supplemental Figure 4**. Locus plots are
834 generated with pyGenomeTracks v3.5 tools (Ramírez et al., 2018).

835 For **Figure 4A**, we compared SNP scores of the effect allele across each model and grouped
836 them by whether they overlapped a cortical or striatal NeuN+ OCR, NeuN- OCR, both, or neither,
837 depending on whether the model was for EXC or D1/D2 cell types, respectively. We computed 2-
838 tailed t-tests between groups and corrected for multiple comparisons with the family-wise
839 Bonferroni method for N=18 tests from three models and (4 choose 2) six possible comparisons
840 per model. * P < 0.05/N, ** P < 0.01/N, *** P < 0.001/N.

841

842 **DATA AVAILABILITY**

843 Code used to run intermediate and final analyses reported in this paper are available on the GitHub
844 page: https://github.com/pfenninglab/addiction_gwas_enrichment.
845 Sequencing output files for data generated in this study are deposited on the GEO at accession

XXXXX. Questions and comments about data and analyses may be directed to the corresponding author.

848 **Bibliography**

849 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM,
850 Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global
851 reference for human genetic variation. Nature 526:68–74.
852 Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of
853 DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33:831–838.
854 Barman P, Reddy D, Bhaumik SR (2019) Mechanisms of Antisense Transcription Initiation with
855 Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis.
856 Noncoding RNA 5.
857 Beaulieu C (1993) Numerical data on neocortical neurons in adult rat, with special reference to
858 the GABA population. Brain Res 609:284–292.
859 Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M (2016) FINEMAP:
860 efficient variable selection using summary data from genome-wide association studies.
861 Bioinformatics 32:1493–1501.
862 Berke JD, Hyman SE (2000) Addiction, dopamine, and the molecular mechanisms of memory.
863 Neuron 25:515–532.
864 Bracci E, Centonze D, Bernardi G, Calabresi P (2002) Dopamine excites fast-spiking
865 interneurons in the striatum. J Neurophysiol 87:2190–2194.
866 Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ (2013) Transposition of native
867 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding
868 proteins and nucleosome position. Nat Methods 10:1213–1218.
869 Buenrostro JD, Wu B, Chang HY, Greenleaf WJ (2015) ATAC-seq: A Method for Assaying
870 Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 109:21.29.1-21.29.9.
871 Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, ReproGen Consortium,
872 Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the
873 Wellcome Trust Case Control Consortium 3, Duncan L, Perry JRB, Patterson N,
874 Robinson EB, Daly MJ, Price AL, Neale BM (2015a) An atlas of genetic correlations
875 across human diseases and traits. Nat Genet 47:1236–1241.
876 Bulik-Sullivan B, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the
877 Psychiatric Genomics Consortium, Patterson N, Daly MJ, Price AL, Neale BM (2015b)
878 LD Score regression distinguishes confounding from polygenicity in genome-wide
879 association studies. Nat Genet 47:291–295.
880 Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol
881 8:e1002822.
882 Cabana-Domínguez J, Shivalikanjli A, Fernàndez-Castillo N, Cormand B (2019) Genome-wide
883 association meta-analysis of cocaine dependence: Shared genetics with comorbid
884 conditions. Prog Neuropsychopharmacol Biol Psychiatry 94:109667.
885 Cannon ME, Currin KW, Young KL, Perrin HJ, Vadlamudi S, Safi A, Song L, Wu Y, Wabitsch
886 M, Laakso M, Crawford GE, Mohlke KL (2019) Open Chromatin Profiling in Adipose
887 Tissue Marks Genomic Regions with Functional Roles in Cardiometabolic Traits. G3
888 (Bethesda) 9:2521–2533.
889 Chan KY, Jang MJ, Yoo BB, Greenbaum A, Ravi N, Wu W-L, Sánchez-Guardado L, Lois C,
890 Mazmanian SK, Deverman BE, Gradinaru V (2017) Engineered AAVs for efficient
891 noninvasive gene delivery to the central and peripheral nervous systems. Nat Neurosci
892 20:1172–1179.
893 Chen L, Fish AE, Capra JA (2018) Prediction of gene regulatory enhancers across species

894 reveals evolutionarily conserved sequence properties. PLoS Comput Biol 14:e1006484.
895 Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, Schaid
896 DJ (2015) Fine Mapping Causal Variants with an Approximate Bayesian Method Using
897 Marginal Test Statistics. Genetics 200:719–736.
898 Cheng Y, Huang CCY, Ma T, Wei X, Wang X, Lu J, Wang J (2017) Distinct synaptic
899 strengthening of the striatal direct and indirect pathways drives alcohol consumption.
900 Biol Psychiatry 81:918–929.
901 Cheng Z, Zhou H, Sherva R, Farrer LA, Kranzler HR, Gelernter J (2018) Genome-wide
902 Association Study Identifies a Regulatory Variant of RGMA Associated With Opioid
903 Dependence in European Americans. Biol Psychiatry 84:762–770.
904 Coffee and Caffeine Genetics Consortium et al. (2015) Genome-wide meta-analysis identifies six
905 novel loci associated with habitual coffee consumption. Mol Psychiatry 20:647–656.
906 Dashti HS et al. (2019) Genome-wide association study identifies genetic loci for self-reported
907 habitual sleep duration supported by accelerometer-derived estimates. Nat Commun
908 10:1100.
909 Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K,
910 Baymuradov UK, Narayanan AK, Onate KC, Graham K, Miyasato SR, Dreszer TR,
911 Strattan JS, Jolanki O, Tanaka FY, Cherry JM (2018) The Encyclopedia of DNA
912 elements (ENCODE): data portal update. Nucleic Acids Res 46:D794–D801.
913 Dick DM (2016) The genetics of addiction: where do we go from here? J Stud Alcohol Drugs
914 77:673–675.
915 Eddie D, Greene MC, White WL, Kelly JF (2019) Medical burden of disease among individuals
916 in recovery from alcohol and other drug problems in the united states: findings from the
917 national recovery survey. J Addict Med 13:385–395.
918 ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the
919 human genome. Nature 489:57–74.
920 Erzurumluoglu AM et al. (2019) Meta-analysis of up to 622,409 individuals identifies 40 novel
921 smoking behaviour associated genetic loci. Mol Psychiatry.
922 Farrell MR, Schoch H, Mahler SV (2018) Modeling cocaine relapse in rodents: Behavioral
923 considerations and circuit mechanisms. Prog Neuropsychopharmacol Biol Psychiatry
924 87:33–47.
925 Fehr C, Yakushev I, Hohmann N, Buchholz H-G, Landvogt C, Deckers H, Eberhardt A, Kläger
926 M, Smolka MN, Scheurich A, Dielentheis T, Schmidt LG, Rösch F, Bartenstein P,
927 Gründer G, Schreckenberger M (2008) Association of low striatal dopamine d2 receptor
928 availability with nicotine dependence similar to that seen with other drugs of abuse. Am J
929 Psychiatry 165:507–514.
930 Ferguson SM, Eskenazi D, Ishikawa M, Wanat MJ, Phillips PEM, Dong Y, Roth BL, Neumaier
931 JF (2011) Transient neuronal inhibition reveals opposing roles of indirect and direct
932 pathways in sensitization. Nat Neurosci 14:22–24.
933 Finucane HK et al. (2015) Partitioning heritability by functional annotation using genome-wide
934 association summary statistics. Nat Genet 47:1228–1235.
935 Finucane HK et al. (2018) Heritability enrichment of specifically expressed genes identifies
936 disease-relevant tissues and cell types. Nat Genet 50:621–629.
937 Fullard JF, Hauberg ME, Bendl J, Egervari G, Cirnaru M-D, Reach SM, Motl J, Ehrlich ME,
938 Hurd YL, Roussos P (2018) An atlas of chromatin accessibility in the adult human brain.
939 Genome Res 28:1243–1252.

940 GBD 2016 Alcohol and Drug Use Collaborators (2018) The global burden of disease attributable
941 to alcohol and drug use in 195 countries and territories, 1990-2016: a systematic analysis
942 for the Global Burden of Disease Study 2016. Lancet Psychiatry 5:987–1012.
943 Ghandi M, Lee D, Mohammad-Noori M, Beer MA (2014) Enhanced regulatory sequence
944 prediction using gapped k-mer features. PLoS Comput Biol 10:e1003711.
945 Gjoneska E, Pfenning AR, Mathys H, Quon G, Kundaje A, Tsai L-H, Kellis M (2015)
946 Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's
947 disease. Nature 518:365–369.
948 Goldstein RZ, Volkow ND (2011) Dysfunction of the prefrontal cortex in addiction:
949 neuroimaging findings and clinical implications. Nat Rev Neurosci 12:652–669.
950 GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–
951 585.
952 GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot
953 analysis: multitissue gene regulation in humans. Science 348:648–660.
954 GTEx Consortium et al. (2017) Genetic effects on gene expression across human tissues. Nature
955 550:204–213.
956 Hickey G, Paten B, Earl D, Zerbino D, Haussler D (2013) HAL: a hierarchical format for storing
957 and analyzing multiple genome alignments. Bioinformatics 29:1341–1342.
958 Howson JMM et al. (2017) Fifteen new risk loci for coronary artery disease highlight arterial-
959 wall-specific mechanisms. Nat Genet 49:1113–1119.
960 International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC)
961 and OCD Collaborative Genetics Association Studies (OCGAS) (2018) Revealing the
962 complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Mol
963 Psychiatry 23:1181–1188.
964 Jensen KP (2016) A Review of Genome-Wide Association Studies of Stimulant and Opioid Use
965 Disorders. Mol Neuropsychiatry 2:37–45.
966 Ji X, Saha S, Kolpakova J, Guildford M, Tapper AR, Martin GE (2017) Dopamine Receptors
967 Differentially Control Binge Alcohol Drinking-Mediated Synaptic Plasticity of the Core
968 Nucleus Accumbens Direct and Indirect Pathways. J Neurosci 37:5463–5474.
969 Jiang C, Wang X, Le Q, Liu P, Liu C, Wang Z, He G, Zheng P, Wang F, Ma L (2019) Morphine
970 coordinates SST and PV interneurons in the prelimbic cortex to disinhibit pyramidal
971 neurons and enhance reward. Mol Psychiatry.
972 Johnston KJA, Adams MJ, Nicholl BI, Ward J, Strawbridge RJ, Ferguson A, McIntosh AM,
973 Bailey MES, Smith DJ (2019) Genome-wide association study of multisite chronic pain
974 in UK Biobank. PLoS Genet 15:e1008164.
975 Kalita CA, Brown CD, Freiman A, Isherwood J, Wen X, Pique-Regi R, Luca F (2018) High-
976 throughput characterization of genetic effects on DNA-protein binding and gene
977 transcription. Genome Res 28:1701–1708.
978 Karlsson Linnér R et al. (2019) Genome-wide association analyses of risk tolerance and risky
979 behaviors in over 1 million individuals identify hundreds of loci and shared genetic
980 influences. Nat Genet 51:245–257.
981 Kelley DR, Reshef YA, Bileschi M, Belanger D, McLean CY, Snoek J (2018) Sequential
982 regulatory activity prediction across chromosomes with convolutional neural networks.
983 Genome Res 28:739–750.
984 Kelley DR, Snoek J, Rinn JL (2016) Basset: learning the regulatory code of the accessible
985 genome with deep convolutional neural networks. Genome Res 26:990–999.

986 Kemp JP et al. (2017) Identification of 153 new loci associated with heel bone mineral density
987 and functional involvement of GPC6 in osteoporosis. Nat Genet 49:1468–1475.
988 Kendler KS, Prescott CA (1998a) Cannabis use, abuse, and dependence in a population-based
989 sample of female twins. Am J Psychiatry 155:1016–1022.
990 Kendler KS, Prescott CA (1998b) Cocaine use, abuse and dependence in a population-based
991 sample of female twins. Br J Psychiatry 173:345–350.
992 Khan A, Riudavets Puig R, Boddie P, Mathelier A (2020) BiasAway: command-line and web
993 server to generate nucleotide composition-matched DNA background sequences.
994 Available at: https://biasaway.uio.no [Accessed July 1, 2020].
995 Kichaev G, Roytman M, Johnson R, Eskin E, Lindström S, Kraft P, Pasaniuc B (2017) Improved
996 methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics 33:248–255.
997 Kim C, Kim S, Lee KY, Kim NH, Kang E-Y, Oh Y-W, Shin C (2019) Differences in bone
998 density on chest CT according to smoking status in males without chronic obstructive
999 lung disease. Sci Rep 9:10467.
1000 Koob GF, Volkow ND (2010) Neurocircuitry of addiction. Neuropsychopharmacology 35:217–
1001 238.
1002 Koob GF, Volkow ND (2016) Neurobiology of addiction: a neurocircuitry analysis. Lancet
1003 Psychiatry 3:760–773.
1004 Krienen FM et al. (2019) Innovations in primate interneuron repertoire. BioRxiv.
1005 Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, Duong TE, Gao D, Chun J, Kharchenko
1006 PV, Zhang K (2018) Integrative single-cell analysis of transcriptional and epigenetic
1007 states in the human adult brain. Nat Biotechnol 36:70–80.
1008 Landt SG et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE
1009 consortia. Genome Res 22:1813–1831.
1010 Lansink CS, Goltstein PM, Lankelma JV, Pennartz CMA (2010) Fast-spiking interneurons of the
1011 rat ventral striatum: temporal coordination of activity with principal cells and
1012 responsiveness to reward. Eur J Neurosci 32:494–508.
1013 Lee D (2016) LS-GKM: a new gkm-SVM for large-scale datasets. Bioinformatics 32:2196–
1014 2198.
1015 Lee IS, Leem AY, Lee SH, Rhee Y, Ha Y, Kim YS (2016) Relationship between pulmonary
1016 function and bone mineral density in the Korean National Health and Nutrition
1017 Examination Survey. Korean J Intern Med 31:899–909.
1018 Lee JH, Ribeiro EA, Kim J, Ko B, Kronman H, Jeong YH, Kim JK, Janak PH, Nestler EJ, Koo
1019 JW, Kim J-H (2020) Dopaminergic regulation of nucleus accumbens cholinergic
1020 interneurons demarcates susceptibility to cocaine addiction. Biol Psychiatry.
1021 Lee JJ et al. (2018) Gene discovery and polygenic prediction from a genome-wide association
1022 study of educational attainment in 1.1 million individuals. Nat Genet 50:1112–1121.
1023 Lefort S, Tomm C, Floyd Sarria JC, Petersen CCH (2009) The excitatory neuronal network of
1024 the C2 barrel column in mouse primary somatosensory cortex. Neuron 61:301–316.
1025 Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for
1026 assigning sequence reads to genomic features. Bioinformatics 30:923–930.
1027 Liu C, Wang M, Wei X, Wu L, Xu J, Dai X, Xia J, Cheng M, Yuan Y, Zhang P, Li J, Feng T,
1028 Chen A, Zhang W, Chen F, Shang Z, Zhang X, Peters BA, Liu L (2019a) An ATAC-seq
1029 atlas of chromatin accessibility in mouse tissues. Sci Data 6:65.
1030 Liu M et al. (2019b) Association studies of up to 1.2 million individuals yield new insights into
1031 the genetic etiology of tobacco and alcohol use. Nat Genet 51:237–244.

1032 Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for
1033 RNA-seq data with DESeq2. Genome Biol 15:550–550.
1034 Melé M et al. (2015) The human transcriptome across tissues and individuals. Science 348:660–
1035 665.
1036 Minnoye L, Taskiran II, Mauduit D, Fazio M, Van Aerschot L, Hulselmans G, Christiaens V,
1037 Makhzami S, Seltenhammer M, Karras P, Primot A, Cadieu E, van Rooijen E, Marine J-
1038 C, Egidy G, Ghanem GE, Zon L, Wouters J, Aerts S (2020) Cross-species analysis of
1039 enhancer logic using deep learning. Genome Res.
1040 Mo A, Mukamel EA, Davis FP, Luo C, Henry GL, Picard S, Urich MA, Nery JR, Sejnowski TJ,
1041 Lister R, Eddy SR, Ecker JR, Nathans J (2015) Epigenomic signatures of neuronal
1042 diversity in the mammalian brain. Neuron 86:1369–1384.
1043 Pasman JA et al. (2018) GWAS of lifetime cannabis use reveals new risk loci, genetic overlap
1044 with psychiatric traits, and a causal influence of schizophrenia. Nat Neurosci 21:1161–
1045 1170.
1046 Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D (2011) Cactus: Algorithms for
1047 genome multiple sequence alignment. Genome Res 21:1512–1528.
1048 Pear VA, Ponicki WR, Gaidus A, Keyes KM, Martins SS, Fink DS, Rivera-Aguirre A,
1049 Gruenewald PJ, Cerdá M (2019) Urban-rural variation in the socioeconomic determinants
1050 of opioid overdose. Drug Alcohol Depend 195:66–73.
1051 Pelechano V, Steinmetz LM (2013) Gene regulation by antisense transcription. Nat Rev Genet
1052 14:880–893.
1053 Pullen E, Oser C (2014) Barriers to substance abuse treatment in rural and urban communities:
1054 counselor perspectives. Subst Use Misuse 49:891–901.
1055 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de
1056 Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association
1057 and population-based linkage analyses. Am J Hum Genet 81:559–575.
1058 Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar
1059 A, Manke T (2018) High-resolution TADs reveal DNA sequences underlying genome
1060 organization in flies. Nat Commun 9:189.
1061 Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F,
1062 Manke T (2016) deepTools2: a next generation web server for deep-sequencing data
1063 analysis. Nucleic Acids Res 44:W160-5.
1064 Ribeiro EA et al. (2018) Transcriptional and physiological adaptations in nucleus accumbens
1065 somatostatin interneurons that regulate behavioral responses to cocaine. Nat Commun
1066 9:3149.
1067 Roadmap Epigenomics Consortium et al. (2015) Integrative analysis of 111 reference human
1068 epigenomes. Nature 518:317–330.
1069 Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, Bien E, Baum M,
1070 Bortolin L, Wang S, Goeva A, Nemesh J, Kamitaki N, Brumbaugh S, Kulp D, McCarroll
1071 SA (2018) Molecular Diversity and Specializations among the Cells of the Adult Mouse
1072 Brain. Cell 174:1015–1030.e16.
1073 Scaplen KM, Kaun KR (2016) Reward from bugs to bipeds: a comparative approach to
1074 understanding how reward circuits function. J Neurogenet 30:133–148.
1075 Schall TA, Wright WJ, Dong Y (2020) Nucleus accumbens fast-spiking interneurons in
1076 motivational and addictive behaviors. Mol Psychiatry.
1077 Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological

1078 insights from 108 schizophrenia-associated genetic loci. Nature 511:421–427.
1079 Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-
1080 wide predictions. Nat Rev Genet 15:272–286.
1081 Smith LN (2018) A disciplined approach to neural network hyper-parameters: Part 1 -- learning
1082 rate, batch size, momentum, and weight decay. arXiv.
1083 Tak YG, Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome
1084 engineering to understand the functional relevance of SNPs in non-coding regions of the
1085 human genome. Epigenetics Chromatin 8:57.
1086 Tepper JM, Koós T (2017) Gabaergic interneurons of the striatum. In: Handbook of basal
1087 ganglia structure and function, second edition, pp 157–178 Handbook of behavioral
1088 neuroscience. Elsevier.
1089 Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS,
1090 Lander ES, Schaffner SF, Sabeti PC (2016) Direct Identification of Hundreds of
1091 Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell 165:1519–
1092 1529.
1093 Thurman RE et al. (2012) The accessible chromatin landscape of the human genome. Nature
1094 489:75–82.
1095 Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL, Reddy
1096 TE (2015) Massively parallel quantification of the regulatory effects of noncoding
1097 genetic variation in a human cohort. Genome Res 25:1206–1214.
1098 Volkow ND, Chang L, Wang G-J, Fowler JS, Ding Y-S, Sedler M, Logan J, Franceschi D,
1099 Gatley J, Hitzemann R, Gifford A, Wong C, Pappas N (2003) Low level of brain
1100 dopamine d_2 receptors in methamphetamine abusers: association with metabolism in the
1101 orbitofrontal cortex. Focus (Madison) 1:150–157.
1102 Volkow ND, Morales M (2015) The brain on drugs: from reward to addiction. Cell 162:712–
1103 725.
1104 Volkow ND, Wang GJ, Fowler JS, Logan J, Gatley SJ, Hitzemann R, Chen AD, Dewey SL,
1105 Pappas N (1997) Decreased striatal dopaminergic responsiveness in detoxified cocaine-
1106 dependent subjects. Nature 386:830–833.
1107 Volkow ND, Wang GJ, Fowler JS, Logan J, Hitzemann R, Ding YS, Pappas N, Shea C, Piscani
1108 K (1996) Decreases in dopamine receptors but not in dopamine transporters in alcoholics.
1109 Alcohol Clin Exp Res 20:1594–1598.
1110 Volkow ND, Wang G-J, Tomasi D, Baler RD (2013) Unbalanced neuronal circuits in addiction.
1111 Curr Opin Neurobiol 23:639–648.
1112 Waaktaar T, Kan K-J, Torgersen S (2018) The genetic and environmental architecture of
1113 substance use development from early adolescence into young adulthood: a longitudinal
1114 twin study of comorbidity of alcohol, tobacco and illicit drug use. Addiction 113:740–
1115 748.
1116 Wang GJ, Volkow ND, Fowler JS, Logan J, Abumrad NN, Hitzemann RJ, Pappas NS, Pascani K
1117 (1997) Dopamine D2 receptor availability in opiate-dependent subjects before and after
1118 naloxone-precipitated withdrawal. Neuropsychopharmacology 16:174–182.
1119 Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, Wu C, Zheng Z, Zhao K, Wang Z, Yi X,
1120 Zhang S, Liu X, Liu Z, Chen K, Yu Y, Sham PC, Li MJ (2020) CAUSALdb: a database
1121 for disease/trait causal variants identified using summary statistics of genome-wide
1122 association studies. Nucleic Acids Res 48:D807–D816.
1123 Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants

1124 from high-throughput sequencing data. Nucleic Acids Res 38:e164.
1125 Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017) Functional mapping and
1126 annotation of genetic associations with FUMA. Nat Commun 8:1826.
1127 Wilar G, Shinoda Y, Sasaoka T, Fukunaga K (2019) Crucial Role of Dopamine D2 Receptor
1128 Signaling in Nicotine-Induced Conditioned Place Preference. Mol Neurobiol 56:7911–
1129 7928.
1130 Wiltschko AB, Pettibone JR, Berke JD (2010) Opposite effects of stimulant and antipsychotic
1131 drugs on striatal fast-spiking interneurons. Neuropsychopharmacology 35:1261–1270.
1132 Worsley Hunt R, Mathelier A, Del Peso L, Wasserman WW (2014) Improving analysis of
1133 transcription factor binding sites within ChIP-Seq data based on topological motif
1134 enrichment. BMC Genomics 15:472.
1135 Xu Z, Liang Q, Song X, Zhang Z, Lindtner S, Li Z, Wen Y, Liu G, Guo T, Qi D, Wang M,
1136 Wang C, Li H, You Y, Wang X, Chen B, Feng H, Rubenstein JL, Yang Z (2018) SP8 and
1137 SP9 coordinately promote D2-type medium spiny neuron production by activating Six3
1138 expression. Development 145.
1139 Yue F et al. (2014) A comparative encyclopedia of DNA elements in the mouse genome. Nature
1140 515:355–364.
1141 Zeng X, Liu D, Zhao X, Chao L, Li Y, Li H, Li W, Gui L, Wu W (2019) Association of bone
1142 mineral density with lung function in a Chinese general population: the Xinxiang rural
1143 cohort study. BMC Pulm Med 19:239.
1144 Zhang X, Kaplow IM, Wirthlin M, Park TY, Pfenning AR (2020) HALPER facilitates the
1145 identification of regulatory element orthologs across species. Bioinformatics.
1146 Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-
1147 based sequence model. Nat Methods 12:931–934.
1148 Zillikens MC et al. (2017) Large meta-analysis of genome-wide association studies identifies
1149 five loci for lean body mass. Nat Commun 8:80.
1150

1151 **Figure 1. Substance use and risky behavior GWAS risk variants enrich within reward**
1152 **region- and cell type-specific epigenomic profiles.**
1153 Partitioned LDSC regression (GWAS enrichment) finds enrichment of substance use and risky
1154 behavior traits in region-specific and cell type-specific open chromatin profiles of human
1155 postmortem brain. **(A)** Pie chart of ANNOVAR-annotated(Wang et al., 2010) SNP function of
1156 addiction-associated trait lead and off-lead SNPs in LD $R^2 > 0.8$. Dark colors indicate un-
1157 transcribed/non-coding annotations, light for transcribed/exonic annotations. SNP annotation
1158 labels are according to ANNOVAR using ENSEMBL build 85 gene annotations (Methods). **(B)**
1159 GWAS enrichment false-discovery rates in ATAC-seq of 14 postmortem human brain regions
1160 coupled with NeuN-labeled fluorescence activated nuclei sorting(Fullard et al., 2018). Brain
1161 regions are stratified by cortical and subcortical regions, with cortical regions ordered frontal to
1162 caudal. Sorted cell types within each brain region are denoted by shape (blue triangle for
1163 NeuN+/neuronal, red circle for NeuN-/glial). FDR-adjustment was performed across all
1164 enrichments on the Fullard *et al.* dataset for Figure 1A and Figure 3: **Cell type-specific machine**
1165 **learning models refine human NeuN+ enrichments for substance use genetic risk GWAS.**
1166 **(A**) Scheme to predict cell type-specific activity of NeuN+ ATAC-seq peaks enriched from brain
1167 regions assayed in Fullard *et al*. (Fullard et al., 2018) using ML models trained on mouse cell-type
1168 specific ATAC-seq peaks. ML-predicted OCRs are input into GWAS enrichment. (**B)** Partitioned
1169 LD score regression of addiction associated traits in Fullard *et al.* NeuN+ OCRs predicted to be
1170 cell type-specific by machine learning models of open chromatin. Cell types are colored by the
1171 source mouse cell type-specific OCRs from **Error! Reference source not found.A**. Original
1172 enrichments from Figure 1A are reproduced in black. Bolded points are significant for FDR < 0.05.

1173

1174 **Figure 4: Machine learning (ML) models for predicting cell type-specific open chromatin**
1175 **predict activity of addiction GWAS SNPs**

1176 **(A)** Cell type-activity prediction scores from the effect allele of genome-wide significant lead
1177 SNPs and off-lead SNPs in LD $R^2 > 0.8$. Activity scores for SNPs are stratified by overlap with
1178 Fullard *et al*. (Fullard et al., 2018) cortical or striatal NeuN+ and NeuN- peaks. Significance
1179 symbols denote Bonferroni-adjusted p-values from 2-tailed t-tests for N=18 possible pairwise
1180 comparisons, N.S. not significant, * $P < 0.05/N$, ** $P < 0.01/N$, *** $P < 0.001/N$. **(B)** Locus plot
1181 candidate SNP with predicted function SNP impact in EXC, D1, and D2 cell types. Genome tracks
1182 from top to bottom: NeuN+ MACS2 ATAC-seq foldchange signal of regions enriched in Figure
1183 1A, SNP tracks showing SNP filtering criteria down to candidate functional SNP (Methods), gene
1184 annotation tracks from GENCODE GRCh38, eQTL link tracks of FDR-significant GTEX cis-
1185 eQTL from cortical and striatal brains, and mouse-human orthologous putative CREs mapped from
1186 cSNAIL ATAC-seq. NeuN+ ATAC-seq tracks and eQTL links are colored by source brain region
1187 as cortical (teal) or striatal (blue). Cell types colors label cortical excitatory neurons (EXC; red),
1188 D1 medium spiny neurons (D1; blue), or D2 medium spiny neurons (D2; green).

1189

1190 **Figure 5 Summary of LDSC GWAS enrichments in human and mouse-human orthologous**
1191 **bulk tissue and cell type open chromatin**

1192 **(A)** Schematic of human NeuN-labeled bulk tissue and occipital cortex cell types from **Figure 1**
1193 for which addiction-associated genetic variants were significantly enriched (FDR < 0.05) in OCRs.
1194 Brain regions are labelled by the cell type that enriched (NeuN+ : blue box/shading; NeuN- : red
1195 box/shading) spatially along with the trait(s) for which OCRs were found significantly enriched

1196 with risk variants. Occipital cortex cell types from **Figure 1C** (same color scheme) are listed along
1197 with the trait(s) for which OCRs were found significantly enriched with risk variants. **(B)**
1198 Schematic of addiction-associated genetic variants that share enrichments from human brain
1199 regions and neuronal subtypes in both human and mouse-human orthologous open chromatin.
1200 Brain graphic adapted from Fullard *et al*.(Fullard et al., 2018)
1201
1202 **Supplemental Figure 1. Shared and unique genetic architecture of genetic risk variants of**
1203 **addiction-associated traits.**
1204 **(A)** LDSC genetic correlation (r_g) matrix of seven addiction-associated traits. FDR-significant
1205 correlations at shown in bold, non-significant in gray (FDR < 0.05). **(B)** Upset plot of genomic
1206 loci shared or unique to each addiction-associated trait. Genomic loci are identified by shared
1207 GWAS-significant SNPs and genomic region overlap.
1208
1209
1210 **Supplemental Figure** . **(C)** GWAS enrichment false-discovery rates in single cell THS-seq OCRs
1211 of major cell clusters in occipital cortex(Lake et al., 2018). Cell types in brain regions that are
1212 enriched at FDR < 0.05 are plotted with bigger shapes and with black outlines. Traits assessed are
1213 age of smoking initiation (AgeofInitiation), average number of cigarettes per day for ever smokers
1214 (CigarettesPerDay), having ever regularly smoked (Smoking Initiation), current versus former
1215 smokers (SmokingCessation), number of alcoholic drinks per week (DrinksPerWeek)(Liu et al.,
1216 2019b), lifetime cannabis use (Cannabis)(Pasman et al., 2018), and risky behavior
1217 (RiskyBehavior)(Karlsson Linnér et al., 2019). OFC: orbitofrontal cortex, VLPFC: ventrolateral
1218 prefrontal cortex, DLPFC: dorsolateral prefrontal cortex, ACC: anterior cingulate cortex, INS:

1219 insula, STC: superior temporal gyrus, ITC: inferior temporal gyrus, PMC: primary motor cortex,
1220 PVC: primary visual cortex, AMY: amygdala, HIPP: hippocampus, MDT: mediodorsal thalamus,
1221 NAc: nucleus accumbens, PUT: putamen, Ast: astrocyte, End: endothelial, Ex: excitatory neuron,
1222 In: inhibitory neuron, Mic: microglia, Oli: oligodendrocyte, Opc: oligodendrocyte precursor.
1223

1224 **Figure 2: Cell type-specific enrichment of substance use traits are conserved in mouse-**
1225 **human orthologous open chromatin regions.**

1226 **(A)** Experimental design to map human orthologous regions from mouse ATAC-seq of bulk cortex
1227 (CTX), dorsal striatum (CPU), and nucleus accumbens (NAc) of cre-dependent Sun1-GFP Nuclear
1228 Anchored Independent Labeled (cSNAIL) nuclei of D1-cre, D2-cre, PV-cre, and SST-cre mice.
1229 cSNAIL ATAC-seq experiments report enriched (+) nuclei populations. **(B)** Partitioned LD score
1230 regression finds enrichment of substance use and risky behavior traits for brain region and cell
1231 type specific ATAC-seq open chromatin profiles of mouse brain. Replication of enrichment is
1232 shown using INTACT-enriched OCRs from Mo *et al*(Mo et al., 2015) of cortical excitatory
1233 (EXC+), vasoactive intestinal peptide interneuron (VIP+), and parvalbumin interneuron (PV+).
1234 Enrichments that are enriched at FDR < 0.05 are plotted with black outlines. FDR-adjusted p-value
1235 was performed across all mouse-human ortholog GWAS enrichment across Figure 2.
1236

1237 **Figure 3: Cell type-specific machine learning models refine human NeuN+ enrichments for**
1238 **substance use genetic risk GWAS.**

1239 **(A**) Scheme to predict cell type-specific activity of NeuN+ ATAC-seq peaks enriched from brain
1240 regions assayed in Fullard *et al*. (Fullard et al., 2018) using ML models trained on mouse cell-type
1241 specific ATAC-seq peaks. ML-predicted OCRs are input into GWAS enrichment. (**B)** Partitioned

1242 LD score regression of addiction associated traits in Fullard *et al.* NeuN+ OCRs predicted to be
1243 cell type-specific by machine learning models of open chromatin. Cell types are colored by the
1244 source mouse cell type-specific OCRs from **Error! Reference source not found.A**. Original
1245 enrichments from Figure 1A are reproduced in black. Bolded points are significant for FDR < 0.05.

1246

1247 **Figure 4: Machine learning (ML) models for predicting cell type-specific open chromatin**
1248 **predict activity of addiction GWAS SNPs**

1249 **(A)** Cell type-activity prediction scores from the effect allele of genome-wide significant lead
1250 SNPs and off-lead SNPs in LD $R^2 > 0.8$. Activity scores for SNPs are stratified by overlap with
1251 Fullard *et al*. (Fullard et al., 2018) cortical or striatal NeuN+ and NeuN- peaks. Significance
1252 symbols denote Bonferroni-adjusted p-values from 2-tailed t-tests for N=18 possible pairwise
1253 comparisons, N.S. not significant, * $P < 0.05/N$, ** $P < 0.01/N$, *** $P < 0.001/N$. **(B)** Locus plot
1254 candidate SNP with predicted function SNP impact in EXC, D1, and D2 cell types. Genome tracks
1255 from top to bottom: NeuN+ MACS2 ATAC-seq foldchange signal of regions enriched in Figure
1256 1A, SNP tracks showing SNP filtering criteria down to candidate functional SNP (Methods), gene
1257 annotation tracks from GENCODE GRCh38, eQTL link tracks of FDR-significant GTEX cis-
1258 eQTL from cortical and striatal brains, and mouse-human orthologous putative CREs mapped from
1259 cSNAIL ATAC-seq. NeuN+ ATAC-seq tracks and eQTL links are colored by source brain region
1260 as cortical (teal) or striatal (blue). Cell types colors label cortical excitatory neurons (EXC; red),
1261 D1 medium spiny neurons (D1; blue), or D2 medium spiny neurons (D2; green).

1262

1263 **Figure 5 Summary of LDSC GWAS enrichments in human and mouse-human orthologous**
1264 **bulk tissue and cell type open chromatin**

1265 **(A)** Schematic of human NeuN-labeled bulk tissue and occipital cortex cell types from **Figure 1**
1266 for which addiction-associated genetic variants were significantly enriched (FDR < 0.05) in OCRs.
1267 Brain regions are labelled by the cell type that enriched (NeuN+ : blue box/shading; NeuN- : red
1268 box/shading) spatially along with the trait(s) for which OCRs were found significantly enriched
1269 with risk variants. Occipital cortex cell types from **Figure 1C** (same color scheme) are listed along
1270 with the trait(s) for which OCRs were found significantly enriched with risk variants. **(B)**
1271 Schematic of addiction-associated genetic variants that share enrichments from human brain
1272 regions and neuronal subtypes in both human and mouse-human orthologous open chromatin.
1273 Brain graphic adapted from Fullard *et al*.(Fullard et al., 2018)

1274

1275 **Supplemental Figure 1. Shared and unique genetic architecture of genetic risk variants of**
1276 **addiction-associated traits.**

1277 **(A)** LDSC genetic correlation (r_g) matrix of seven addiction-associated traits. FDR-significant
1278 correlations at shown in bold, non-significant in gray (FDR < 0.05). **(B)** Upset plot of genomic
1279 loci shared or unique to each addiction-associated trait. Genomic loci are identified by shared
1280 GWAS-significant SNPs and genomic region overlap.

1281

1282

1283 **Supplemental Figure 2. Sensitivity of partitioned LDSC regression for cell type- and region-**
1284 **specific in the GWAS trait enrichment requires well-powered GWAS in relevant cell types.**

1285 GWAS enrichment plots with false-discovery rates in ATAC-seq of 14 postmortem human brain
1286 regions coupled with NeuN-labeled fluorescence activated nuclei sorting(Fullard et al., 2018).
1287 Regions are stratified by cortical and subcortical regions, with cortical regions ordered frontal to
1288 caudal. Sorted cell types within each brain region are denoted by shape (blue triangle for
1289 NeuN+/neuronal, red circle for NeuN-/glial). Cell types in brain regions that are enriched at FDR
1290 < 0.05 are plotted with bigger shapes and with black outlines. **(A)** GWAS enrichment of addiction-
1291 or substance use-associated traits: multi-site chronic pain (ChronicPain)(Johnston et al., 2019),
1292 cocaine dependence (CocaineDep)(Cabana-Domínguez et al., 2019) , opioid dependence
1293 (OpioidDep)(Cheng et al., 2018), diagnosis of obsessive-compulsive disorder (OCD)(International
1294 Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD
1295 Collaborative Genetics Association Studies (OCGAS), 2018), and cups of coffee drank per day
1296 (CoffeePerDay)(Coffee and Caffeine Genetics Consortium et al., 2015). The GWAS for OCD,
1297 opioid dependence, and cocaine dependence are reportedly underpowered to detect genetic
1298 liability for these traits ($N_{case} < 5{,}000$). **(B)** GWAS enrichment in well-powered brain-related traits
1299 show cell type- and region-specific enrichment: educational attainment (EduAttain)(Lee et al.,
1300 2018), schizophrenia risk (Schizophrenia)(Schizophrenia Working Group of the Psychiatric
1301 Genomics Consortium, 2014), habitual sleep duration (SleepDuration)(Dashti et al., 2019). **(C)**
1302 GWAS enrichment in non-brain associated traits do not show cell type- or region-specific
1303 enrichment: heel bone-mineral density (BMD)(Kemp et al., 2017), coronary artery disease
1304 (CAD)(Howson et al., 2017), and lean body mass (LBM)(Zillikens et al., 2017).

1305

1306 **Supplemental Figure 3. Cell type specificity of cSNAIL ATAC-seq in mouse cortex and**
1307 **striatum**

1308 (A) Principle component plots of chromatin accessibility counts from cre-dependent Sun1-GFP
1309 Nuclear Anchored Independent Labeled (cSNAIL) ATAC-seq from *cre*-driver lines (**Methods**).
1310 Major axes of variation separate cell types by tissue source (PC1) and cell type versus bulk ATAC-
1311 seq (PC2). (B) Normalized coverage track plots around marker genes demarcating cell type-
1312 specificity of cSNAIL ATAC-seq samples. (C) Density correlation plot of normalized chromatin
1313 accessibility log counts around the transcription start site (TSS) with matched pseudo-bulk cell
1314 type log gene counts from Drop-seq of mouse cortex and striatum(Saunders et al., 2018). Drop-
1315 seq cell types meta-gene profiles report sum gene counts for cell clusters from frontal cortex and
1316 striatum. Pearson's and Spearman's correlation are denoted with R and ρ, respectively. (D)
1317 Pairwise correlation matrix of TSS chromatin accessibility log counts with Drop-seq pseudo-bulk
1318 log gene counts from cortical and striatal cell clusters.

1319
1320

1321 **Supplemental Figure 4. GWAS enrichment in addiction- and non-addiction-related traits**
1322 **using mapped mouse orthologs of tissue- and cell type-specific open chromatin regions.**
1323 GWAS enrichment plots with false-discovery rates in human orthologous regions mapped from
1324 mouse ATAC-seq of bulk cortex (CTX), dorsal striatum (CPU), and nucleus accumbens (Nac) or
1325 cre-dependent Sun1-GFP Nuclear Anchored Independent Labeled (cSNAIL) nuclei of D1-cre, D2-
1326 cre, and PV-cre mice. cSNAIL ATAC-seq experiments report both enriched (+) and de-enriched
1327 (-) nuclei populations. Enrichments that are enriched at FDR < 0.05 are plot with black outlines.
1328 Replication of enrichment is shown using INTACT-enriched OCRs from Mo *et al*(Mo et al., 2015)
1329 of cortical excitatory (EXC+), vasoactive intestinal peptide interneuron (VIP+), and parvalbumin
1330 interneuron (PV+). **(A)** GWAS enrichment of addiction- or substance use-associated traits: multi-
1331 site chronic pain (`ChronicPain`), cocaine dependence (`CocaineDep`), opioid dependence

1332 (OpioidDep), diagnosis of obsessive-compulsive disorder (OCD), and cups of coffee drank per
1333 day (CoffeePerDay). The GWAS for OCD, opioid dependence, and cocaine dependence are
1334 reportedly underpowered to detect genetic liability for these traits (N_{case}< 5,000). **(B)** GWAS
1335 enrichment in well-powered brain-related traits show cell type- and region-specific enrichment:
1336 educational attainment (EduAttain), schizophrenia risk (Schizophrenia), habitual sleep
1337 duration (SleepDuration). **(C)** GWAS enrichment in non-brain associated traits do not show
1338 cell type- or region-specific enrichment: heel bone-mineral density (BMD), coronary artery disease
1339 (CAD), and lean body mass (LBM). **(D)** Heatmap of LDSC regression coefficients of GWAS
1340 enrichment for all measured GWAS in non-brain OCRs from human or mouse-human mapped
1341 orthologs. Tissues for which OCRs are significantly enriched (FDR < 0.05) with GWAS variants
1342 are outlined with a bolded box.

1343

1344 **Supplemental Figure 5. ML model performance and selection of candidate functional SNPs.**
1345 **(A)** Performance metrics for machine learning (ML) models evaluated on the test sets of IDR peaks
1346 or 10x nucleotide-distribution matched negatives. Five-fold cross validation scheme was used to
1347 train on IDR peaks of each cell type and tissue. Performance metrics are reported for accuracy
1348 and F1-score (using threshold = 0.5) and area under the receiver-operator-characteristic (auROC)
1349 and the precision-recall curve (auPRC) (**Methods**). Models were trained on IDR peaks of mouse
1350 cortical excitatory and striatal D1 and D2 medium spiny neurons. **(B)** Selection criteria from
1351 addiction-associated trait SNPs to candidate SNPs with functional impact in gene-regulatory
1352 function.

1353

1354 **Supplemental Figure 6. Locus plots of addiction-associated SNPs predicted to act in striatal**
1355 **and cortical cell types.**
1356 **(A-D)** Locus plot of candidate SNP with predicted function SNP impact in D1 or D2 cell types.
1357 Genome tracks from top to bottom: NeuN+ MACS2 ATAC-seq foldchange signal of regions
1358 enriched in Figure 1A, SNP tracks showing SNP filtering criteria down to candidate functional
1359 SNP (Methods), gene annotation tracks from GENCODE GRCh38, eQTL link tracks of FDR-
1360 significant GTEX cis-eQTL from cortical and striatal brains, and mouse-human orthologous
1361 putative CREs mapped from cSNAIL ATAC-seq. NeuN+ ATAC-seq tracks and eQTL links are
1362 colored by source brain region as cortical (teal) or striatal (blue). Cell types colors label cortical
1363 excitatory neurons (EXC; red), D1 medium spiny neurons (D1; blue), D2 medium spiny neurons
1364 (D2; green). Panels A and C report SNPs that are known eQTLs in striatal brain tissues. (A)
1365 rs11191352 (SmokingInitiation), (B) rs10742814 (DrinksPerWeek), (C) rs9844736
1366 (RiskyBehavior, SmokingInitiation), (D) Two SNPs ~400bp apart, rs6870603 and rs7712167
1367 (SmokingInitiation).

1368

1369 **Supplemental Table 1. Addiction-associated genetic variants annotated with cell type and**
1370 **brain region functional markers**
1371 Addiction-associated genetic variants from the seven GWAS explored in this study further
1372 annotated by FUMA(Watanabe et al., 2017), CAUSALdb(Wang et al., 2020), overlap with NeuN+
1373 OCRs (Fullard et al., 2018), and ML model open chromatin prediction of the risk and non-risk
1374 alleles as well as ML model delta SNP scores. Several tabs report FUMA annotations, CAUSAL
1375 db annotation, SNP overlap with Fullard et al. NeuN+/NeuN- peaks, CNN scores for effect allele,
1376 CNN scores for non-effect allele, CNN delta SNP scores. All tabs contain unique identifying

1377 columns according to FUMA outputs. Additional columns report HaploReg annotations from these
1378 SNPs where available data exists. The SNP positions are mapped to report hg38 coordinates and
1379 GenomicLocus report overlapping loci across. SNP overlaps with NeuN peaks are indicated as 0
1380 for no overlap and 1 for overlap within IDR peaks. CNN scores are reported average across all
1381 folds and tissues from the same cell type and tissues. CNN scores range from 0 to 1, with scores
1382 larger than 0.5 are predicted active in that cell type. Delta SNP scores, instead, are normally
1383 distributed and centered at 0.

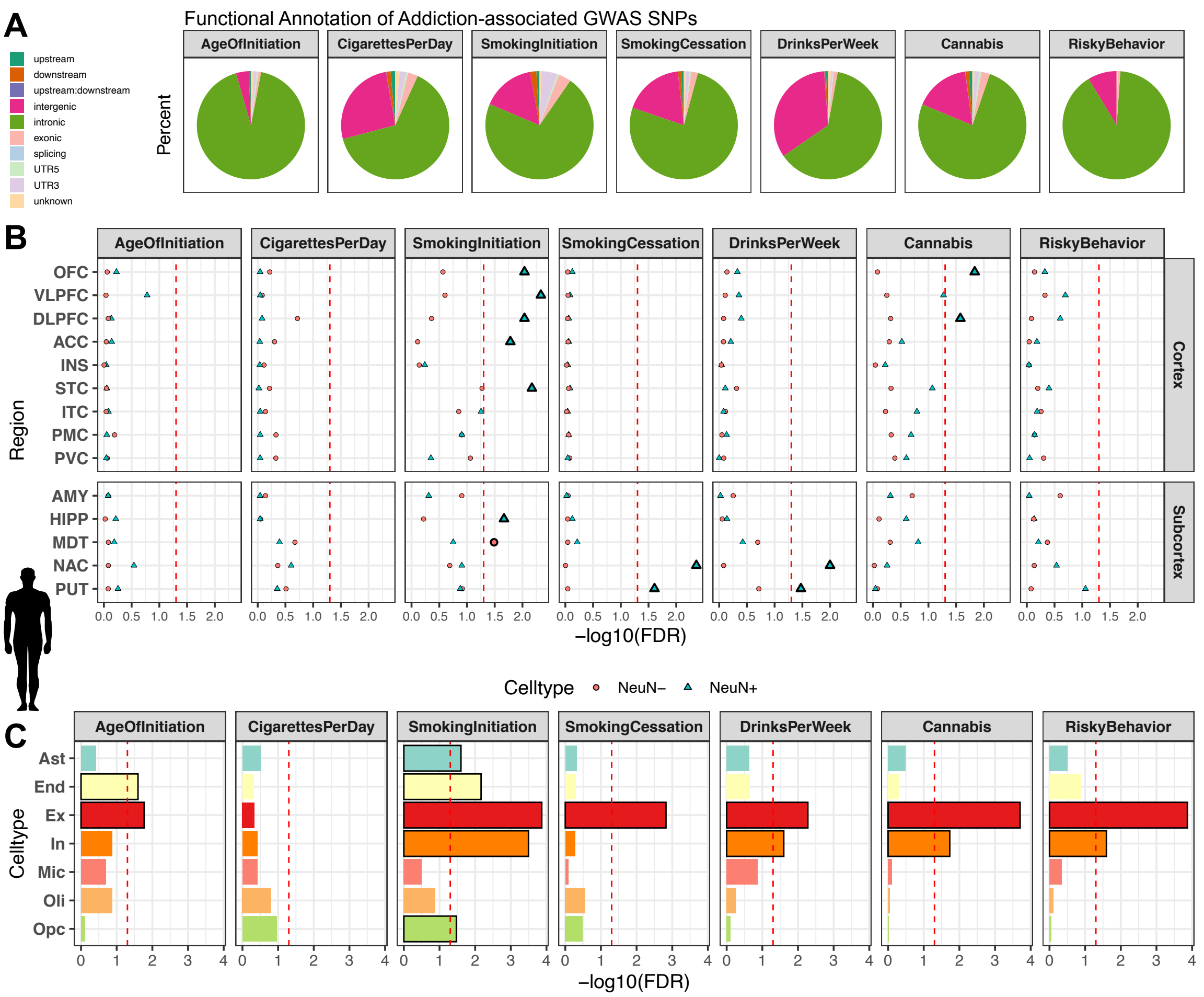

A
Functional Annotation of Addiction-associated GWAS SNPs
upstream
downstream
upstream:downstream
intergenic
intronic
exonic
splicing
UTR5
UTR3
unknown
Percent
AgeOfInitiation
CigarettesPerDay
SmokingInitiation
SmokingCessation
DrinksPerWeek
Cannabis
RiskyBehavior
B
Region
OFC
VLPFC
DLPFC
ACC
INS
STC
ITC
PMC
PVC
Cortex
AMY
HIPP
MDT
NAC
PUT
Subcortex
–log10(FDR)
Celltype
NeuN–
NeuN+
C
Celltype
Ast
End
Ex
In
Mic
Oli
Opc
–log10(FDR)

Figure 1

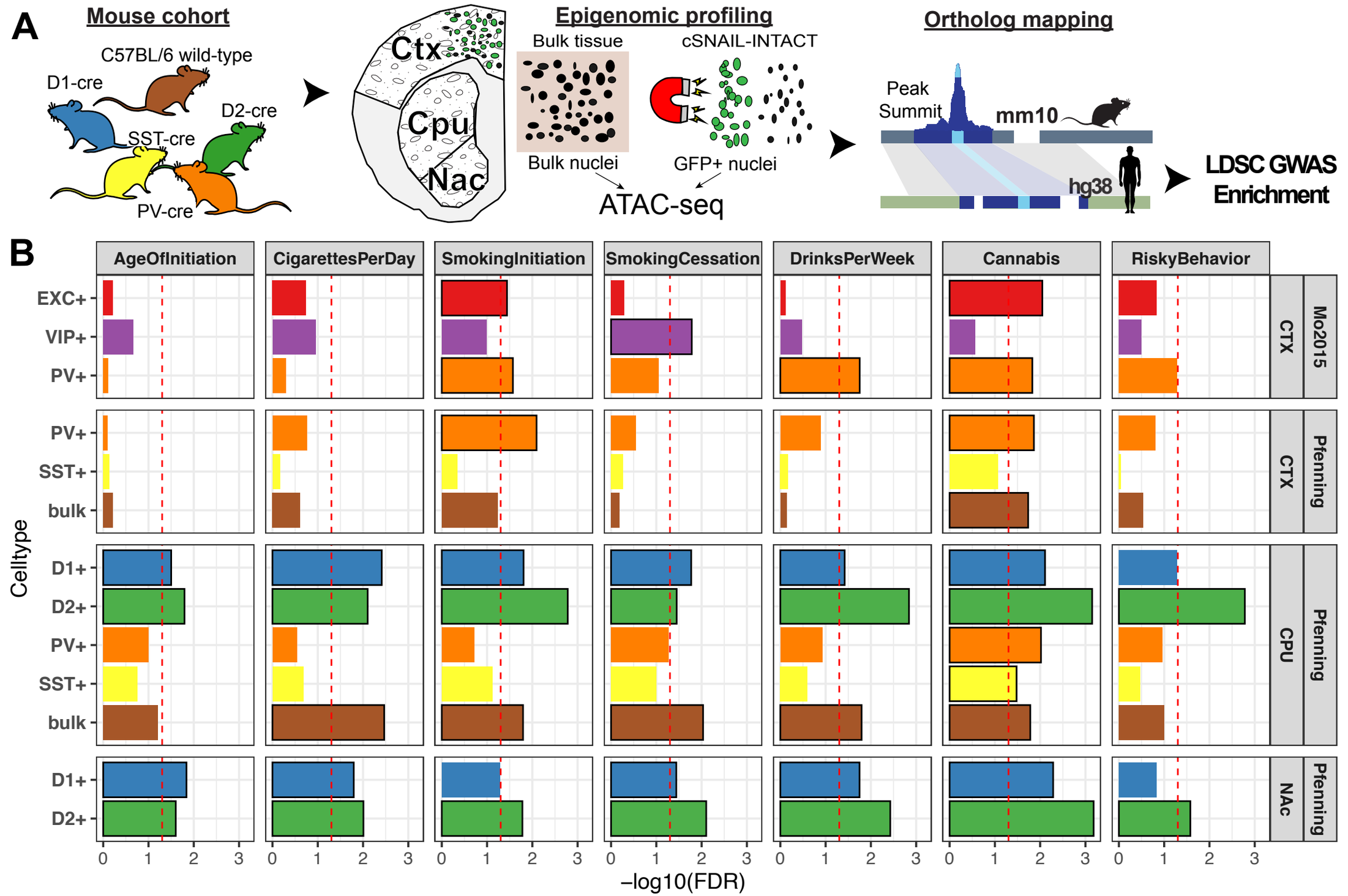

A
Mouse cohort
C57BL/6 wild-type
D1-cre
D2-cre
SST-cre
PV-cre
Ctx
Cpu
Nac
Epigenomic profiling
Bulk tissue
cSNAIL-INTACT
Bulk nuclei
GFP+ nuclei
ATAC-seq
Ortholog mapping
Peak Summit
mm10
hg38
LDSC GWAS Enrichment
B
AgeOfInitiation
CigarettesPerDay
SmokingInitiation
SmokingCessation
DrinksPerWeek
Cannabis
RiskyBehavior
Celltype
EXC+
VIP+
PV+
SST+
bulk
D1+
D2+
CTX
CPU
NAc
Mo2015
Pfenning
–log10(FDR)

Figure 2

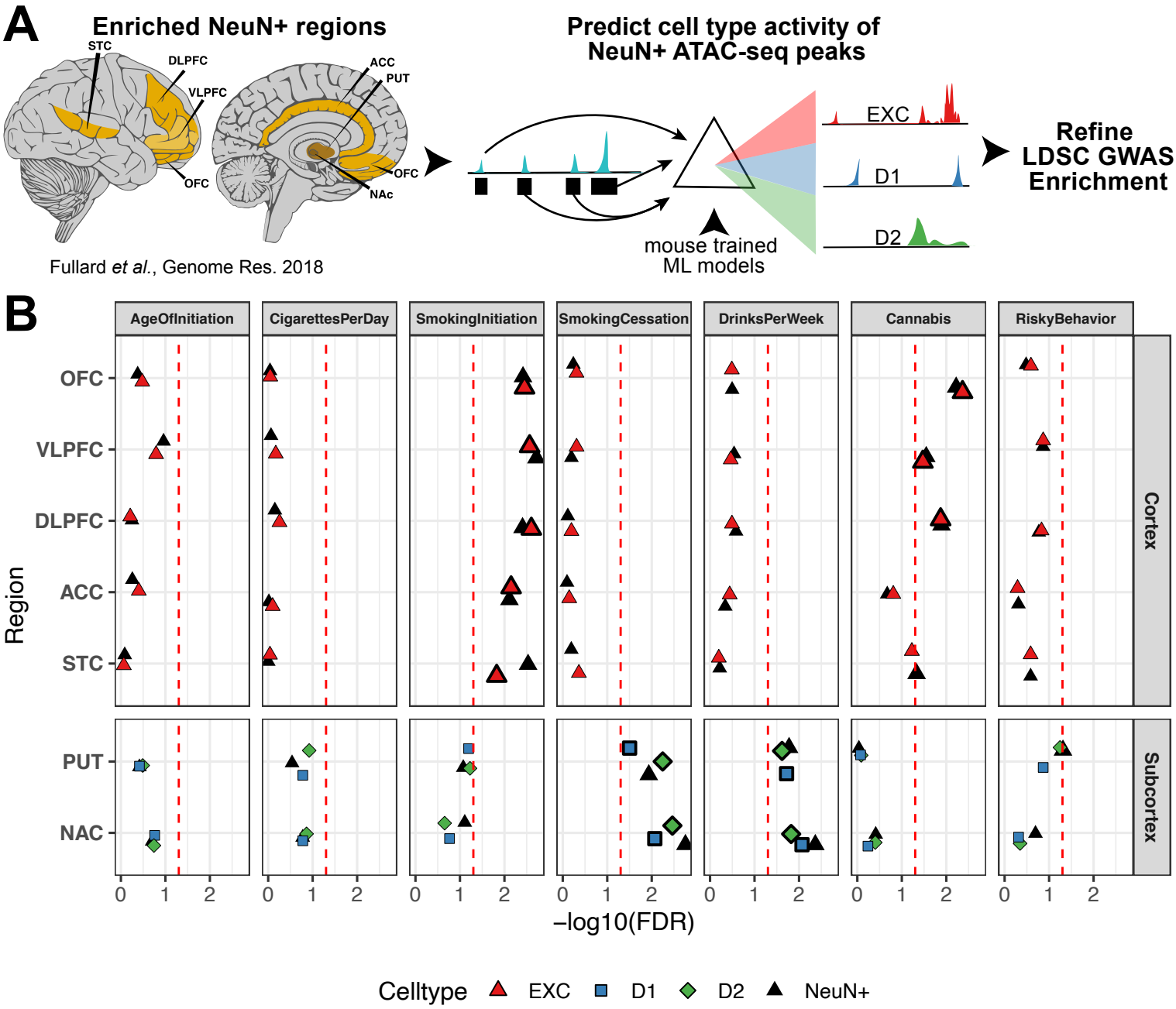

A
Enriched NeuN+ regions
STC
DLPFC
VLPFC
OFC
ACC
PUT
OFC
NAc
Fullard et al., Genome Res. 2018
Predict cell type activity of NeuN+ ATAC-seq peaks
mouse trained ML models
EXC
D1
D2
Refine LDSC GWAS Enrichment
B
AgeOfInitiation
CigarettesPerDay
SmokingInitiation
SmokingCessation
DrinksPerWeek
Cannabis
RiskyBehavior
Region
OFC
VLPFC
DLPFC
ACC
STC
PUT
NAC
Cortex
Subcortex
0 1 2
−log10(FDR)
Celltype
EXC
D1
D2
NeuN+

Figure 3

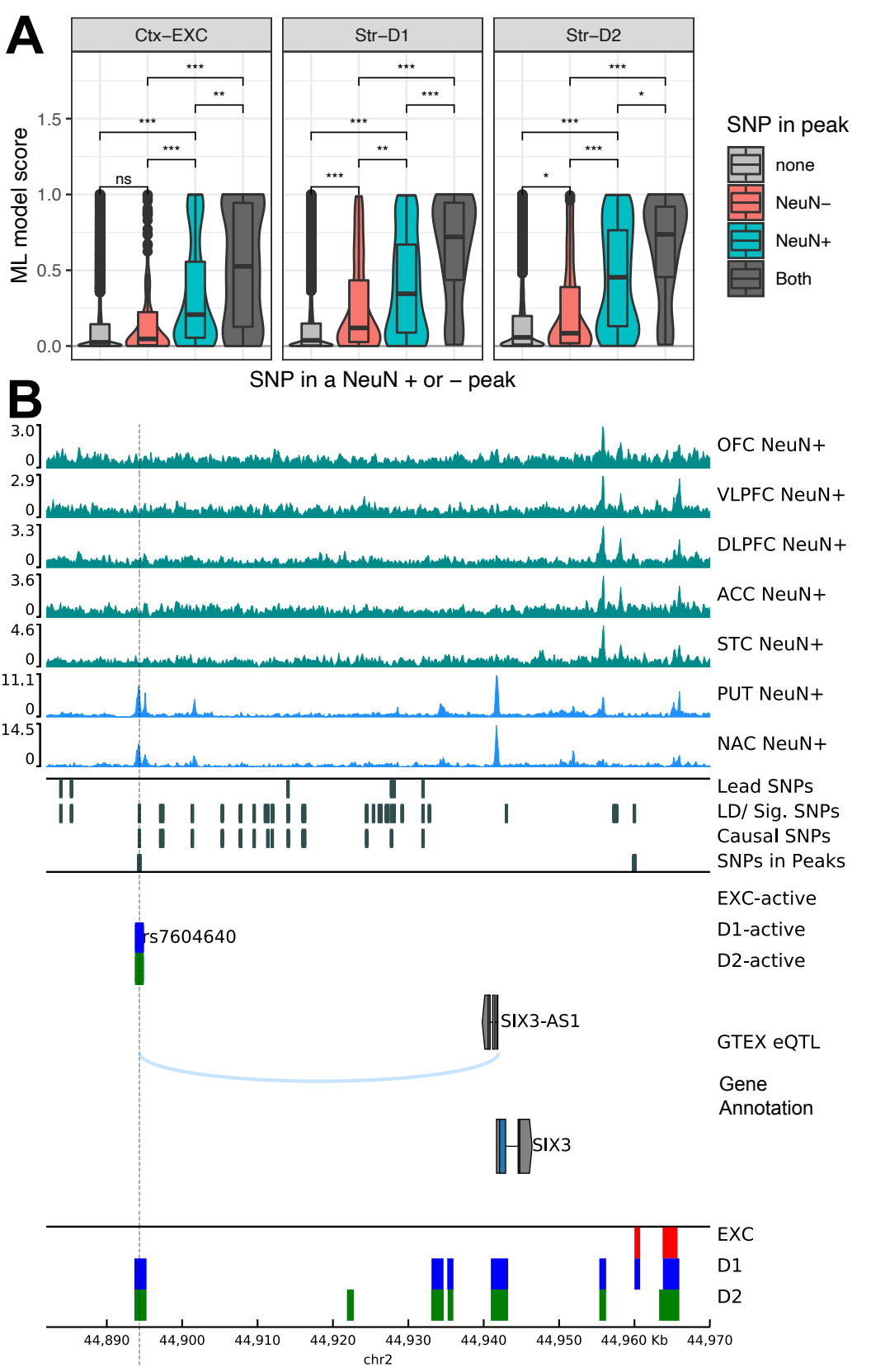

A
Ctx−EXC
Str−D1
Str−D2
ML model score
1.5
1.0
0.5
0.0
ns
***
**
*
SNP in peak
none
NeuN−
NeuN+
Both
SNP in a NeuN + or − peak
B
3.0
0
2.9
3.3
3.6
4.6
11.1
14.5
OFC NeuN+
VLPFC NeuN+
DLPFC NeuN+
ACC NeuN+
STC NeuN+
PUT NeuN+
NAC NeuN+
Lead SNPs
LD/ Sig. SNPs
Causal SNPs
SNPs in Peaks
EXC-active
D1-active
D2-active
rs7604640
SIX3-AS1
GTEX eQTL
Gene Annotation
SIX3
EXC
D1
D2
44,890
44,900
44,910
44,920
44,930
44,940
44,950
44,960 Kb
44,970
chr2

Figure 4

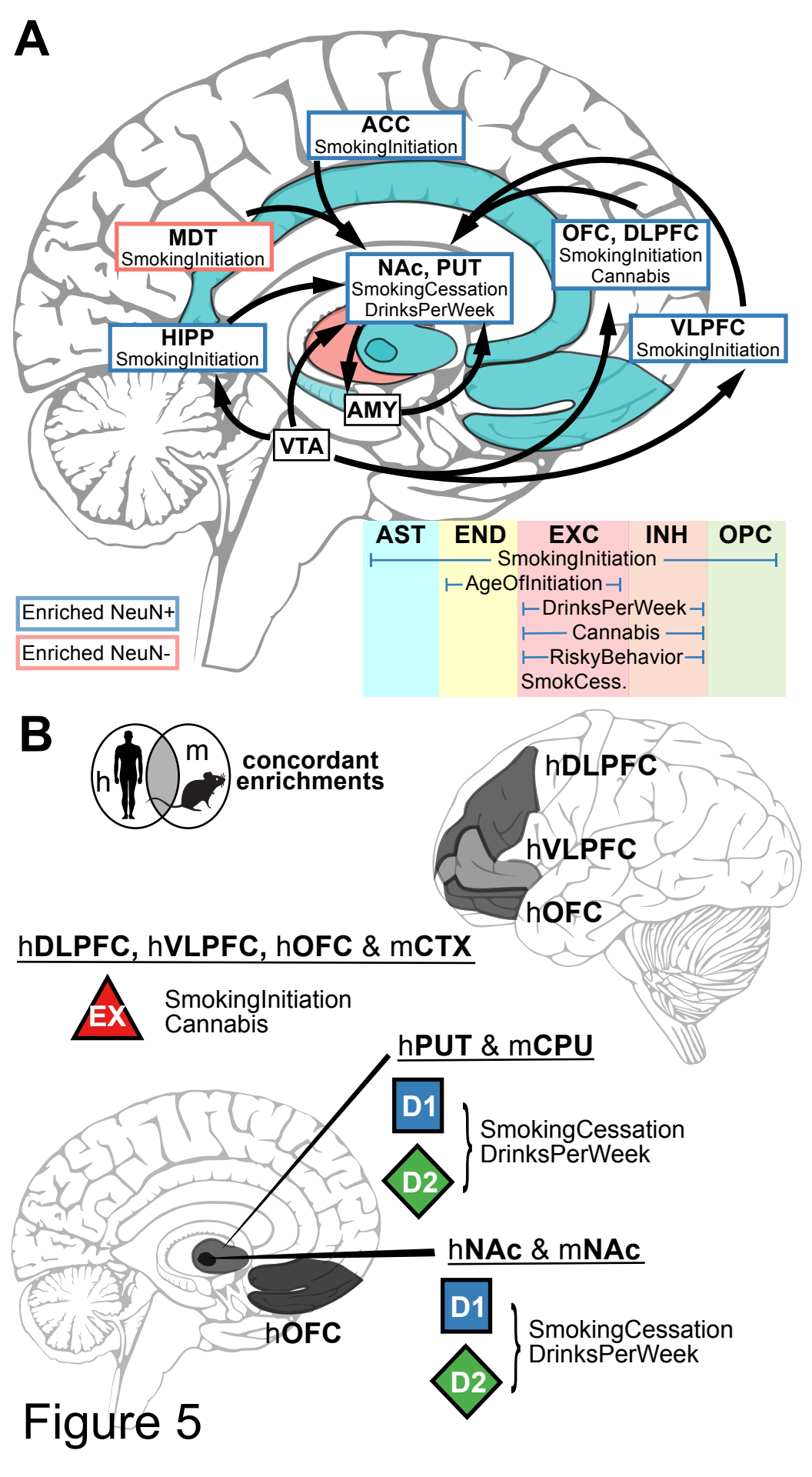
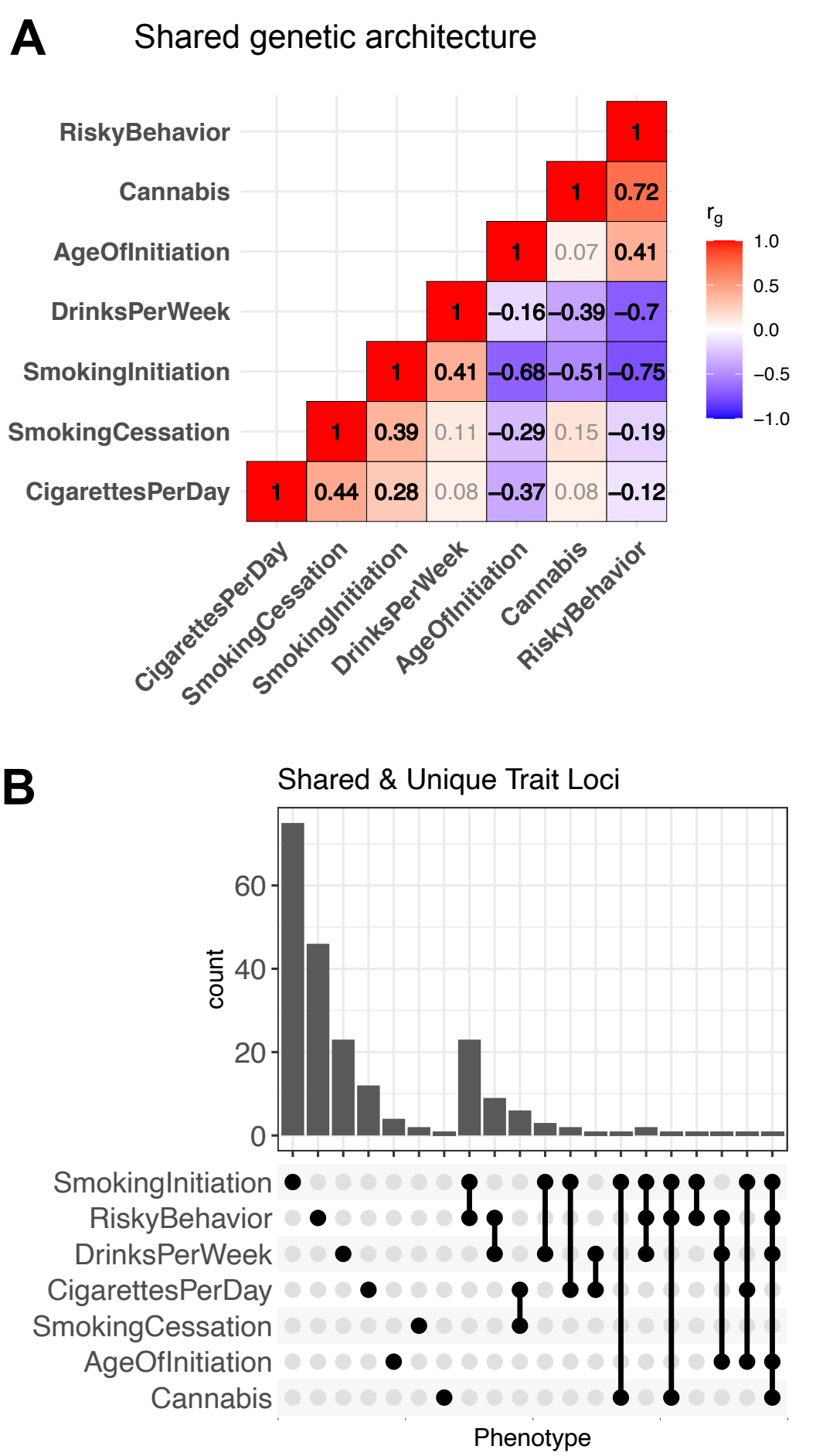
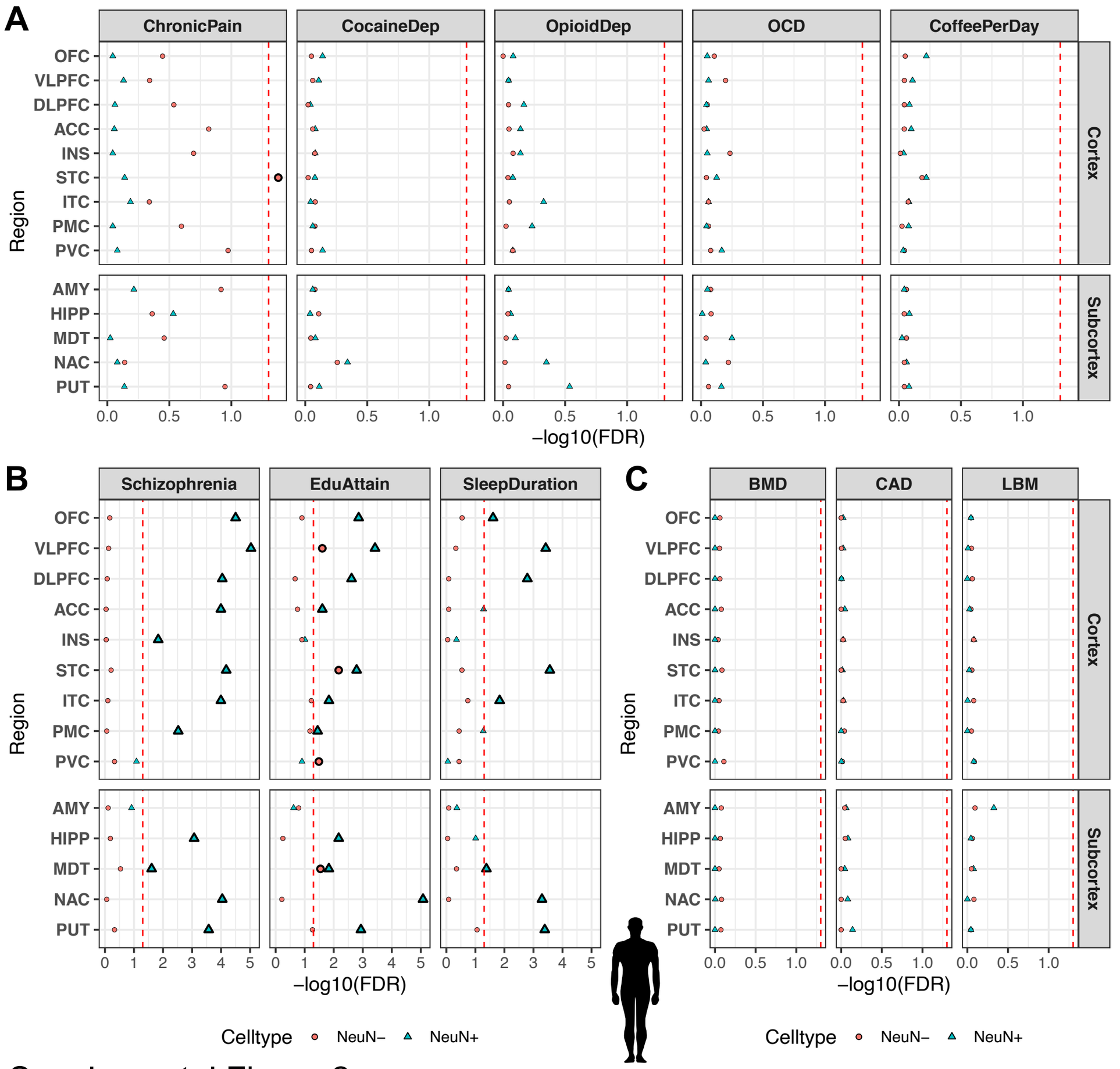

A
ACC
SmokingInitiation
MDT
SmokingInitiation
NAc, PUT
SmokingCessation
DrinksPerWeek
OFC, DLPFC
SmokingInitiation
Cannabis
HIPP
SmokingInitiation
VLPFC
SmokingInitiation
AMY
VTA
AST
END
EXC
INH
OPC
SmokingInitiation
AgeOfInitiation
DrinksPerWeek
Cannabis
RiskyBehavior
SmokCess.
Enriched NeuN+
Enriched NeuN-
B
h
m
concordant enrichments
hDLPFC
hVLPFC
hOFC
hDLPFC, hVLPFC, hOFC & mCTX
EX
SmokingInitiation
Cannabis
hPUT & mCPU
D1
D2
SmokingCessation
DrinksPerWeek
hNAc & mNAc
D1
D2
SmokingCessation
DrinksPerWeek
hOFC

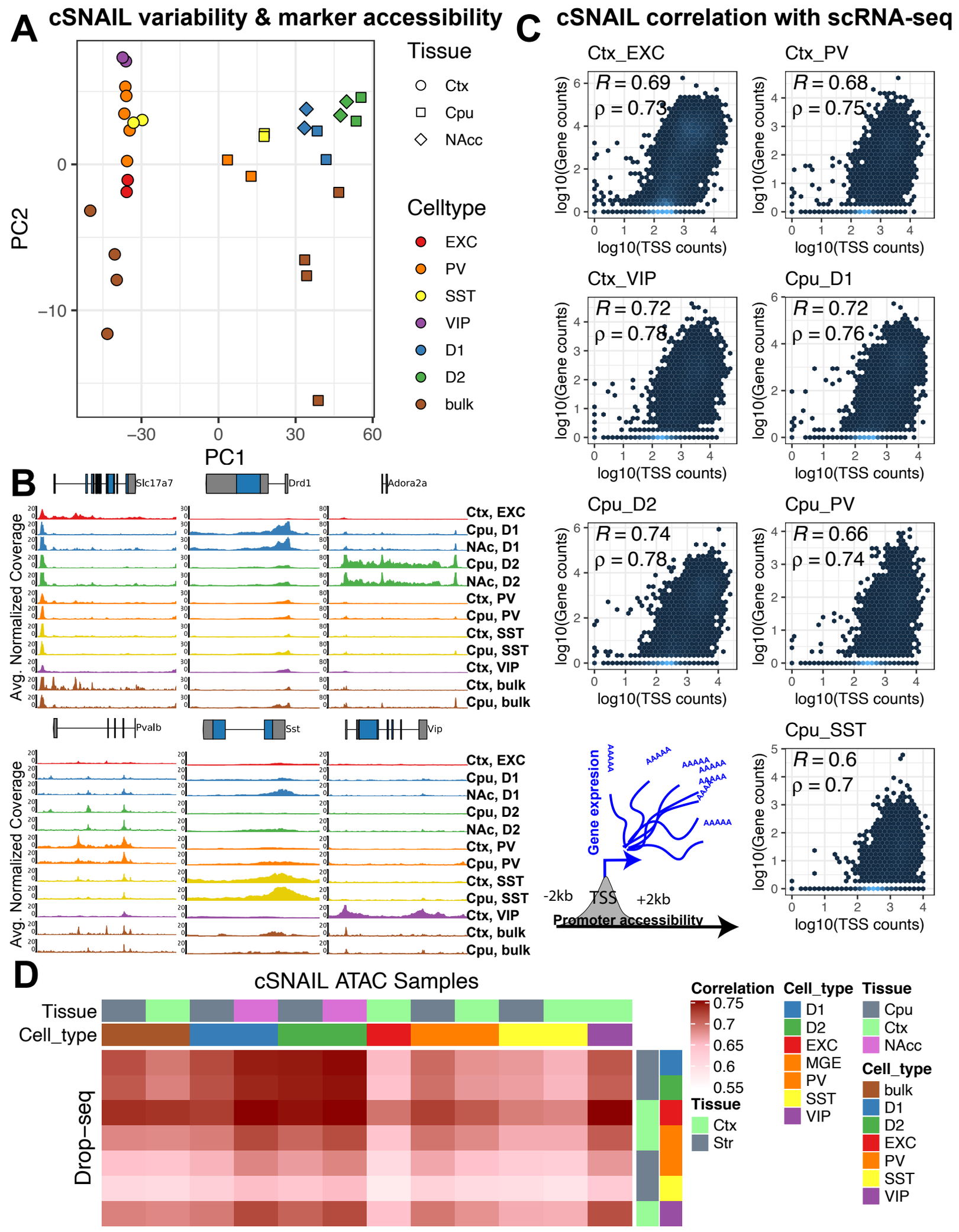
Figure 5



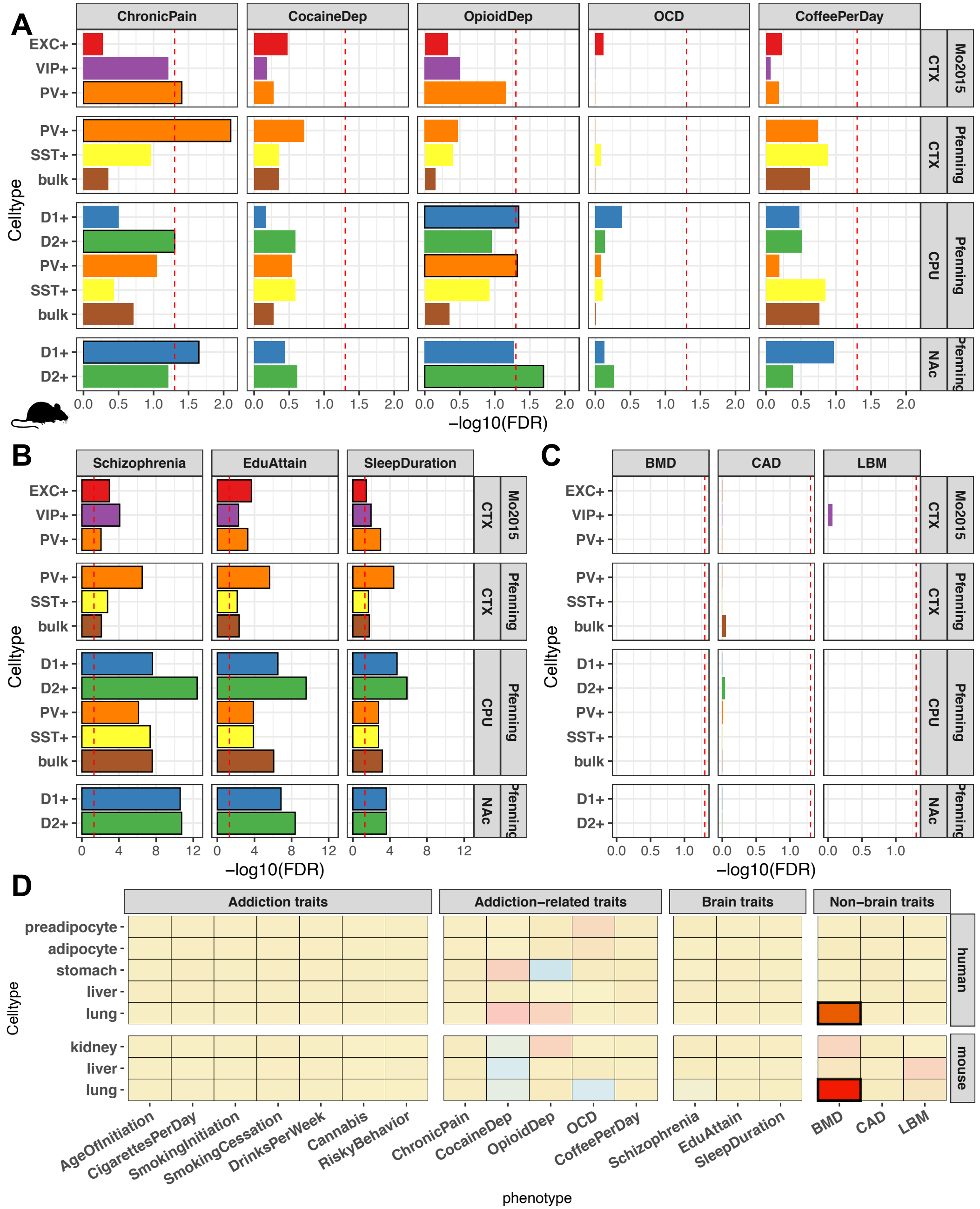
Supplemental Figure 1



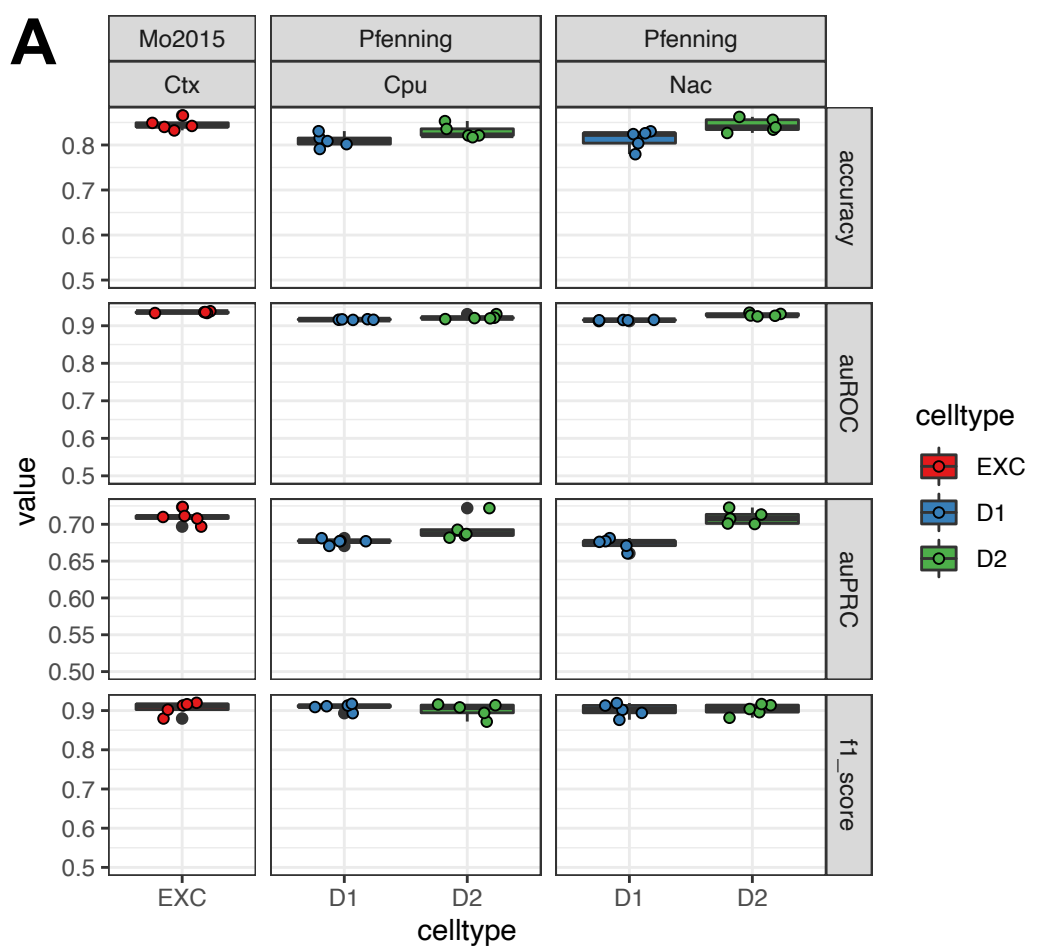
Supplemental Figure 2

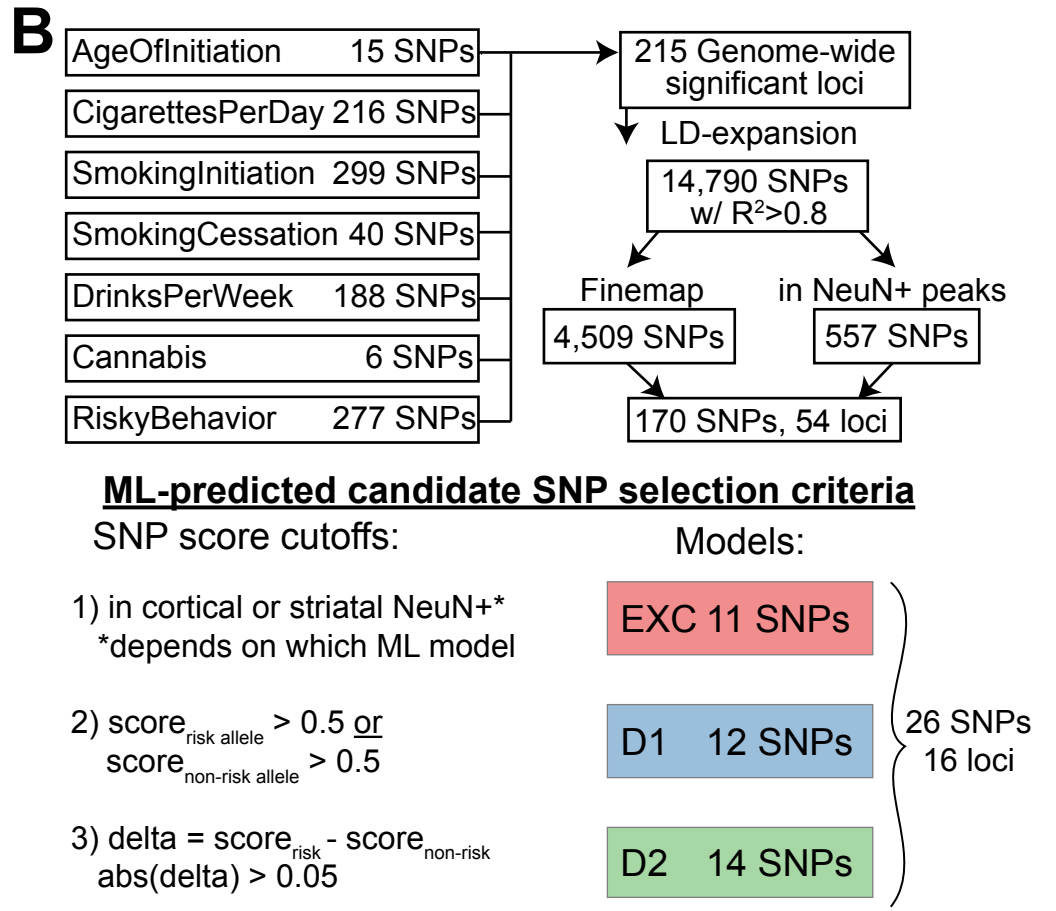


Supplemental Figure 3

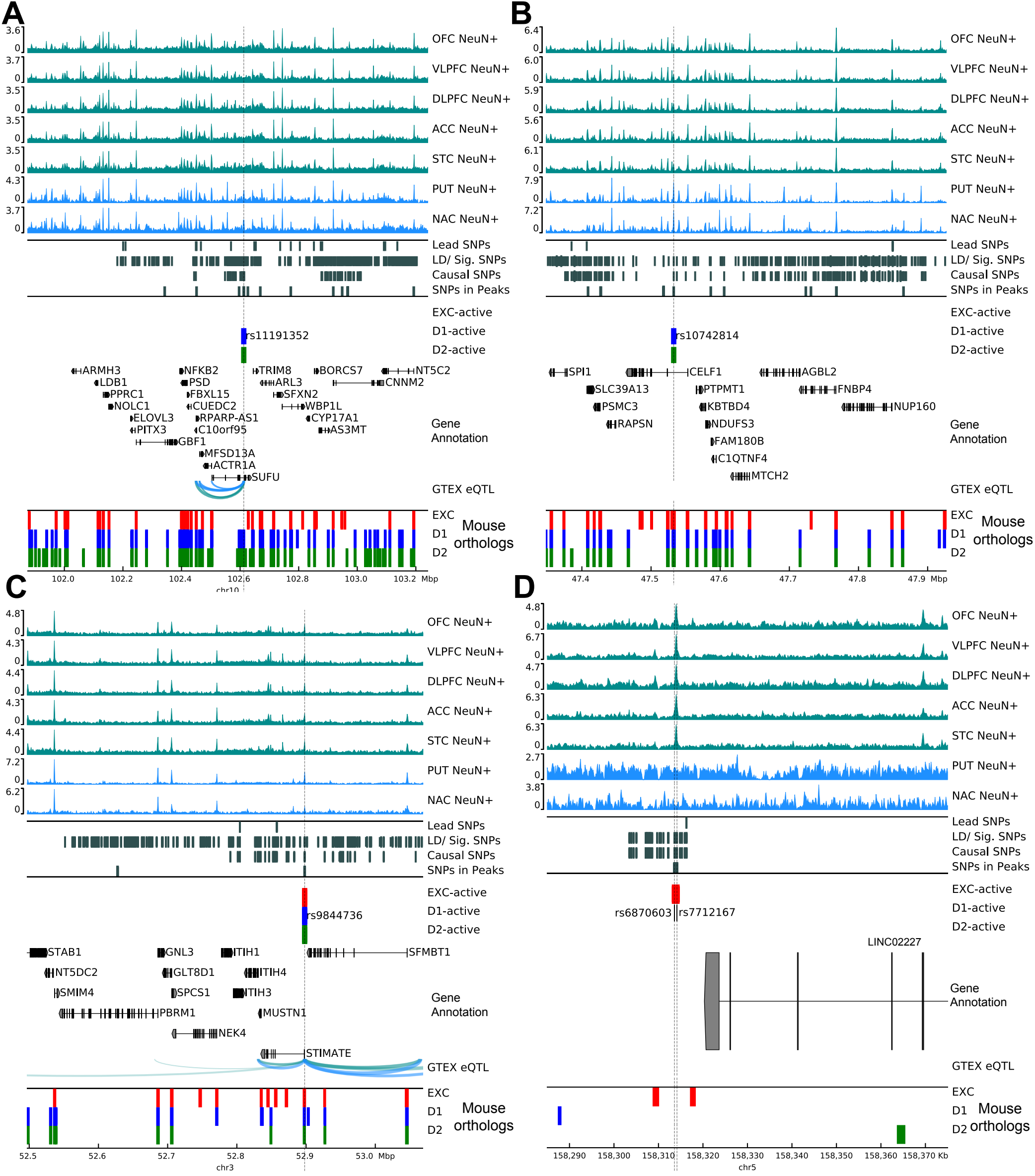


Supplemental Figure 4





Supplemental Figure 5



Supplemental Figure 6