Epigenome-wide analysis uncovers a blood-based DNA methylation biomarker of lifetime cannabis use.

Christina A. Markunas[1*], Dana B. Hancock[1], Zongli Xu[2], Bryan Quach[1], Dale P. Sandler[2], Eric O. Johnson[1,3,†], and Jack A. Taylor[2,4,†]

[1]Center for Omics Discovery and Epidemiology, RTI International, Research Triangle Park, NC, USA; [2]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA; [3]Fellow Program, RTI International, Research Triangle Park, NC, USA; [4]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA

*Corresponding author:

Christina A. Markunas, PhD
3040 East Cornwallis Rd, PO Box 12194
Research Triangle Park, NC 27709
Email: cmarkunas@rti.org
Telephone: 1-919-541-5837

[†]*Contributed equally*

ABSTRACT

Background: Cannabis, the most commonly used illicit drug in the US, is associated with adverse and beneficial effects. To better understand the full spectrum of health consequences, biomarkers that accurately classify cannabis use are needed. DNA methylation (DNAm) is an excellent candidate, yet no blood-based epigenome-wide association studies (EWAS) in humans exist.

Methods: We conducted an EWAS using blood-based DNAm data from a case-cohort study within Sister Study, a prospective cohort of women at risk of developing breast cancer (Discovery N=1,730; Replication N=853). We ran robust linear regression models of DNAm by lifetime cannabis (ever versus never) use, adjusting for age, breast cancer, tobacco and alcohol use, technical factors, and cell-type proportions. A multi-CpG classifier of lifetime cannabis use was developed using penalized regression of top EWAS CpGs in the discovery sample and validated in the replication sample. We also examined published *cis*-methylation quantitative trait loci (*cis*-meQTLs) for reported cannabis-associated SNPs in relation to our EWAS findings.

Results: We identified and replicated an association with lifetime cannabis use at cg15973234 (*CEMIP*): combined $P=3.3\times10^{-8}$. We found no overlap between published cannabis-associated SNPs ($P<0.05$) and blood-based *cis*-meQTLs of cg15973234. A 3-CpG classifier of lifetime cannabis use had an area under the curve of 0.79 ($P=2.2\times10^{-16}$) in the replication sample.

Conclusions: Our findings provide evidence that blood-based DNAm is associated with lifetime cannabis use. DNAm levels at cg15973234 were not associated with reported genetic risk factors for lifetime cannabis use, suggesting that the observed DNAm difference was driven by cannabis exposure.

INTRODUCTION

Cannabis use is highly prevalent with 45% of Americans aged 12 or older reporting lifetime use (defined here as ever-use of cannabis), and 15% reporting use in the past year (1). These numbers are expected to grow due to increasing legalization of both medical and recreational cannabis use. As of 2019, 33 US States have legalized cannabis for medical purposes, 10 of which have also legalized recreational use (2). Cannabis use has been associated with various adverse effects, including impaired cognitive and motor function, altered judgement, paranoia, psychosis, cannabis use disorder, altered brain development, chronic bronchitis, and diminished life satisfaction and achievement (3). In contrast, evidence exists supporting therapeutic benefits for various clinical conditions (e.g., glaucoma, acquired immune deficiency syndrome, nausea, chronic pain, inflammation) (3). Efforts to better understand the full spectrum of cannabis-related health consequences have been hindered as a result of under-reporting (e.g., due to social stigma associated with use) and the absence of available biomarkers that can accurately quantify cannabis usage patterns.

Currently available biomarkers of cannabis exposure, such as urinary metabolites, suffer limitations. These existing biomarkers have limited windows for detection, are largely restricted to acute exposures (e.g., ranging from 3 to >30 days, depending on

the frequency of cannabis use (4)), and are unable to quantify cumulative exposure, which may prove to be a better indicator of subsequent health outcomes (4).

Epigenetic modifications represent promising candidates for biomarker research. DNA methylation (DNAm), the most commonly studied form of epigenetic modification, is defined by the presence of a methyl group ($-CH_3$) most often at the carbon-5 position of a cytosine nucleotide in the context of CpG sites (adjacent cytosine and guanine nucleotides linked by a phosphate group). DNAm can be influenced by genetic factors, disease, environmental exposures, and lifestyle, and can vary across developmental stages of life (e.g., infancy, childhood, and adulthood), tissues, and cell types. An important feature of exposure-related DNAm changes is that they can either be persistent (i.e., stable changes) or reversible (i.e., return to prior state) once the exposure is no longer present. This combination of both persistent and reversible changes has value for biomarker development and has been observed, for example, in relation to tobacco smoking where only a subset of DNAm changes identified between current versus never smokers are also found between former versus never smokers (5, 6).

Research geared towards understanding epigenetic changes related to cannabinoids, which encompass endocannabinoids (endogenous ligands), natural cannabinoids (derived from cannabis and includes Δ^9-tetrahydrocannabinol [THC] and cannabidiol), and synthetic cannabinoids, is growing. However, studies to date are largely based on animal studies and human candidate gene studies (7, 8), with only one epigenome-wide study in human sperm (9). To begin to address this gap in the field, we report the first blood-based EWAS of lifetime cannabis (ever vs. never) use with

discovery (N=1,730) and replication (N=853) samples. We extended our EWAS findings by using genetic information to help distinguish genetically- versus exposure-driven effects. We further leveraged our EWAS results to develop the first multi-CpG biomarker predictive of lifetime cannabis (ever vs. never) use.

METHODS AND MATERIALS

Study population

The Sister Study is a prospective cohort of 50,884 women ascertained across the US between 2003–2009 that was designed to examine environmental and genetic risk factors of breast cancer (10). Women, aged 35–75 years old, were enrolled if they had a sister diagnosed with breast cancer and no personal history of breast cancer at baseline. Whole blood samples were collected, along with data from questionnaires and interviews covering demographics, lifestyle, family and medical history, and environmental exposures.

Since the study's inception, a number of sub-studies have been designed to address specific hypotheses related to women's health. For the current study, we used existing DNAm array data from a case-cohort study of 2,878 non-Hispanic white women designed to identify blood-based DNAm changes associated with incident breast cancer, as described previously (11, 12). Briefly, the case-cohort study included a random sample of 1,336 women from the cohort and 1,542 additional women who later developed *in situ* or invasive breast cancer during follow-up (between enrollment and sampling of the case-cohort in 2015). Out of the 2,878 women, 1,616 women (1,542 + 46 cases from the random sample) were diagnosed with breast cancer during follow-up

and 1,262 women (1,336 – 46 cases from the random sample) remained clinically breast cancer free.

Informed written consent was obtained from all Sister Study participants. The Sister Study was approved by the Institutional Review Boards (IRBs) of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health, and the Copernicus Group (http://www.cgirb.com/irb-services/). The current study was approved by the IRB at RTI International.

Cannabis assessment

Lifetime cannabis use was determined by response to the question, "*Have you ever smoked marijuana?*", asked along with questions about tobacco smoking during the baseline computer-assisted telephone interview. Basing our primary analysis on ever-use facilitated our evaluation of genetic data using the largest genome-wide association study (GWAS) of cannabis use reported to date (N = 184,765) which focused on ever-use only (13). Age of initiation, duration of use, and frequency of use were also assessed in the Sister Study and were considered for secondary analyses in the current study. Sister Study questionnaires can be accessed at https://www.sisterstudystars.org/.

DNAm data, quality control (QC), and pre-processing

Blood sample collection and DNA extraction have been described previously (11, 12). A total of 2,878 blood samples were assayed using the Illumina HumanMethylation450 array which covers >450,000 CpG sites targeting promoters, CpG islands, 5' and 3' untranslated regions, the major histocompatibility complex, and some enhancer regions (14).

Data quality assessment and pre-processing was conducted using the R package, *ENmix* (15). A series of diagnostic plots were generated to detect problematic samples, arrays, and laboratory plates. The ENmix pipeline was implemented for data pre-processing and included the following steps: background correction using the ENmix method (15); correction of fluorescent dye-bias using the RELIC method (16); inter-array quantile normalization; and correction of probe type bias using the RCP (regression on correlated CpGs) method (17). Surrogate variables (SVs) of the array control probes were generated to adjust for technical artifacts. To control for cellular heterogeneity, blood cell type proportions (CD8T cells, CD4T cells, natural killer cells [NK], B cell, monocytes [Mono], and granulocytes [Gran]) were estimated following the Houseman method (18). β-values, corresponding to the ratio of methylated intensities relative to the total intensity, were calculated to represent DNAm levels at each CpG.

Both sample- and probe-level exclusions were applied. In total, 295 samples were excluded due to poor bisulfite conversion efficiency (average bisulfite intensity <4000), outlier based on QC diagnostic plots (e.g., DNAm β-value distribution), low call rate (> 5% low quality data [Illumina detection $P>1\times10^{-6}$, number of beads <3, or values outside 3*Interquartile range (IQR)]), related individuals (one sister from each pair was selected at random for exclusion), missing phenotype data, or date of breast cancer diagnosis preceding blood draw. A total of 67,564 probes were excluded: 16,100 probes with > 5% low quality data; 14,522 probes with a common SNP (minor allele frequency >0.01) at the single base extension site or CpG site; 26,799 cross-reactive probes mapping to multiple genomic locations (19); 10,143 probes mapping to sex chromosomes. In addition, extreme DNAm β-value outliers were removed (outside

3*IQR); missing values were imputed using K-nearest neighbor. Following exclusions, the final analysis dataset included 2,583 women and 417,948 CpGs.

EWAS analysis

The Sister Study breast cancer case-cohort sample was randomly divided into discovery (2/3 of the overall sample: N=1,730 [855 ever users]) and replication (1/3 of the overall sample: N=853 [392 ever users]). Characteristics of the case-cohort sample used for analysis are provided in Tables 1 and S1.

For the EWAS, robust linear regression models were implemented using the R package, *MASS (20)*, to test the association between lifetime cannabis use (ever [used at least once in lifetime] versus never use) and methylation (β-value) at each CpG site, adjusting for age (continuous), incident breast cancer status (event, non-event), tobacco smoking (never, former, current), alcohol use (noncurrent, current), laboratory plate, DNA extraction method, six control SVs, and six blood cell type proportions. To adjust for multiple testing, the false discovery rate (FDR) was controlled at 10% (Benjamini and Hochberg) (21). Replication analyses were conducted using the same set of covariates. For replication, a Bonferroni correction was applied accounting for the number of CpGs tested.

A series of sensitivity analyses were conducted to assess other possible confounders. We considered the following: current perceived stress (derived from items 2, 6, 7, and 14 from the 14-item Perceived Stress Scale instrument (22); scale 0–20), family income while growing up (low income, middle income and well-off), body mass index (BMI) as calculated from height and weight measured by examiners at baseline (continuous), and self-reported history of depression (yes, no). To further rule out

possible effects due to an association with subsequent breast cancer and use of alcohol and/or tobacco, we conducted stratified analyses by these variables. EPISTRUCTURE was implemented using the python toolset, GLINT (23), to rule out genetic confounding for significant EWAS findings. More specifically, the first four EPISTRUCTURE components (principal components) were computed, while accounting for estimated cell type proportions, and included as covariates in the model to adjust for population stratification.

Biomarker development and validation

We developed a multi-CpG classifier of lifetime cannabis use within the EWAS discovery sample (N = 1,730) and validated it using the replication sample (N = 853; withheld from model training). Least absolute shrinkage and selection operator (LASSO) regression (24) was used for model training and variable selection, adjusting for the same set of covariates as before. Instead of using all CpGs as a test bed for model training, we selected subsets of CpGs aiming to reduce the amount of noise introduced in the model. As no gold-standard selection criteria exists, two different significance thresholds ($P<1\times10^{-5}$ and $P<1\times10^{-4}$) from our discovery lifetime cannabis use EWAS were applied to select CpGs for input into LASSO regression. The R package, *glmnet* (25), was used to implement LASSO regression and 10-fold cross validation for model selection (based on lambda) in the discovery sample. Model performance was evaluated in the independent replication sample using the R package, *pROC* (26), to perform a receiver operating characteristic (ROC) analysis. We calculated the 95% confidence interval of the area under the ROC curve (AUC) using 5,000 bootstrap iterations in the replication sample.

Follow-up analyses of replicable EWAS findings

To further characterize replicable EWAS findings, we examined the relationship between DNAm levels and total duration of use, divided into quartiles given the skewed distribution (upper [> 5 years of use] versus lower quartile [≤ 1 year of use]; mean +/- SD= 4.59 ± 7.2 years) among ever users in the combined sample (N = 1,239 [N = 8 with missing information]). We also evaluated the effect of age of initiation (log transformed) on DNAm levels (N = 1,245 [N = 2 with missing information]; mean +/- SD: 20.85 ± 6.76 years). These secondary evaluations were restricted to these two exposure characteristics given the high degree of missing information (23%) and limited variability for frequency of use among ever users.

Our EWAS findings could be due to underlying genetic effects or the effect of exposure to cannabis. First, to explore whether genetic risk factors for lifetime cannabis use could explain our findings, we used the set of eight genome-wide significant single nucleotide polymorphisms (SNPs) identified in the most recent lifetime cannabis use GWAS meta-analysis (N = 184,765; European ancestry) (13) and the BIOS QTL browser (27) to identify CpGs associated with the cannabis use-associated SNPs in blood samples (i.e., determine if lifetime cannabis use-associated SNPs are *cis*-methylation quantitative trait loci [*cis*-meQTLs] in blood). We then examined whether any of the CpGs we identified in our EWAS overlapped with these *cis*-meQTL-CpGs. Of note, *cis*-meQTLs provided through the BIOS QTL browser were identified using data on 3,841 Dutch individuals and modeled without accounting for any phenotype or exposure.

Next, we used the BIOS QTL browser (27) to identify blood-based *cis*-meQTLs for the lifetime cannabis-use associated CpGs observed in our EWAS. The Sister Study participants have been genotyped using the OncoArray that is customized to optimally capture cancer-associated SNPs (28), but the 230,000 tagging SNP backbone does not provide sufficient coverage to directly model *cis*-meQTL associations with lifetime cannabis use in the Sister Study case-cohort sample. As such, we again used results from the lifetime cannabis use GWAS meta-analysis (N=162,082 with 23andMe samples removed) (13) and performed a look-up of the association between BIOS-implicated *cis*-meQTLs and lifetime cannabis use.

RESULTS

EWAS sample characteristics

On average, women who reported lifetime cannabis (ever) use were younger and more likely to report alcohol and tobacco use (Table 1). The prevalence of current tobacco smoking was low (7% of the total sample) and, among ever cannabis users, only half identified as a former tobacco smoker. Alcohol use was more prevalent in the sample with 10% of ever cannabis users reporting noncurrent use and only 3 individuals indicating never use. In addition, ever cannabis users were more likely to grow up in a family with higher income levels, have a lower BMI, and have a higher estimated proportion of CD8 T cells and a lower proportion of NK cells (Table S1). However, after controlling for age, the estimated cell type proportions were not associated (P>0.05) with lifetime cannabis use. There was also suggestive evidence in the discovery

sample, but not in the replication sample, for differences between ever and never users by self-reported history of depression and current perceived stress (Table S1).

EWAS of lifetime cannabis use: Discovery and Replication

We identified one epigenome-wide significant association (FDR<0.10) at cg15973234 (Figure 1 and Table S2 [EWAS results with unadjusted P<0.05]). Overall, effects sizes were small (Figure S1), and there was no evidence of inflation (λ=1.02; Figure S2). Further adjustment for potential confounders, including current perceived stress, family income while growing up, BMI, and history of depression did not substantively affect the overall results, thus the more parsimonious model remained our primary model (see Figures S3–S4 for model comparisons).

The cg15973234-lifetime cannabis use association met a Bonferroni correction in the replication sample ($P_{Replication}$=0.04, $\beta_{Replication}$=-0.005; Table 2) and in the combined sample ($P_{Combined}=3.3\times10^{-8}$, $\beta_{Combined}$=-0.008; Table 2). In further testing of cg15973234 in the combined sample, we found that it was not associated with total duration of use (P=0.85; β=0.0005) or age of initiation (P=0.35; β=0.01).

Additional sensitivity analyses designed to further evaluate possible effects of cigarette smoking, alcohol use, incident breast cancer, and genetic ancestry on our observed cg15973234 association did not suggest significant confounding (Table S3).

The lifetime cannabis use-associated cg15973234 lies within a CpG island spanning the 5' untranslated region of the cell migration inducing hyaluronidase 1 (*CEMIP*) gene. The adjacent CpGs in the island are located within 1kb of cg15973234 but are not highly correlated with cg15973234 and thus provide minimal support for the EWAS signal (Table S4). The most highly correlated CpG within the region

(cg24159335; Pearson correlation [r] with cg15973234=0.29) has a similar mean DNAm β-value and is nominally associated with lifetime cannabis use with a direction of effect consistent with cg15973234 ($P=3.2\times10^{-3}$; β=-0.003; Table S4).

Blood-based DNAm biomarker of lifetime cannabis use

Using the top 5 most significant CpGs (EWAS discovery $P<1\times10^{-5}$) as input into LASSO regression for model training resulted in a classifier composed of three CpGs: cg15973234 (*CEMIP*), cg04685163 (*DLGAP2*), and cg03765885 (intergenic) (Table S5). Validation of the model in the independent replication sample produced an AUC=0.79 (95% confidence interval [0.76, 0.82]; $P<2.2\times10^{-16}$ as compared to a random classifier [AUC=0.50] (29); Figure 2). The inclusion of an additional 54 CpGs (Discovery EWAS: $1\times10^{-5}<P<1\times10^{-4}$) for model training in the discovery sample resulted in a 41-CpG classifier (Table S6), including the 3 CpGs from the initial model. The additional 38 CpGs did not improve model performance (AUC in replication sample=0.78 [0.75, 0.81]).

Follow-up of replicable EWAS finding

None of the eight reported lifetime cannabis use-associated SNPs, from the largest GWAS to date (N=184,765) (13), are *cis*-meQTLs of cg15973234 as they are not located on the same chromosome as our EWAS finding. To further assess the possibility that there may be other, weaker genetic risk factors of cannabis use driving our EWAS finding, we used the BIOS QTL browser and identified two independent *cis*-meQTLs for cg15973234 in blood (rs3848177 and rs8025670; FDR<0.05), as determined by statistical modeling of *cis*-meQTL effects (27). To assess the genetic association between these common SNPs and lifetime cannabis use, we again used

results from the lifetime cannabis use GWAS meta-analysis (13) and found no association (Table 3).

DISCUSSION

We report the first blood-based EWAS of cannabis use where we identified and replicated a difference between ever and never users in DNAm levels at cg15973234 (*CEMIP*). DNAm at this site is not correlated with reported cannabis-associated SNPs (13), suggesting that our finding is unlikely to be genetically-driven and more likely related to the cannabis exposure. Further characterization of the cg15973234-lifetime cannabis use association indicated that the finding appears robust to total duration of use and age of initiation, suggesting that this CpG within *CEMIP* acts as a general indicator of lifetime cannabis use (i.e., a marker of ever versus never use).

CEMIP plays a role in hyaluronan binding and degradation (30) and has been implicated in nonsyndromic hearing loss (31), autoimmune disorders (30, 32), and various cancers (33-35). Hyaluronan, one of the main components of the extracellular matrix, is an important regulator of inflammation (36) and immune processes (37). Although cannabinoids are thought to play a role in immunoregulation and have anti-inflammatory properties (38), there is little evidence to date that those processes are modulated by CEMIP or downstream genes in the CEMIP pathway (39, 40).

Our work identifies the first multi-CpG DNAm biomarker of cannabis use, demonstrating in an independent replication sample that lifetime cannabis use can be correctly categorized using DNAm at three CpGs, including our top signal at cg15973234 within *CEMIP* (AUC=0.79). The cannabis-associated CpG with the largest

(magnitude) LASSO coefficient in the model (cg04685163) is located 10 base pairs away from a *DLGAP2* (DLG associated protein 2) transcription start site. Notably, *DLGAP2* is a post-synaptic density protein which has been associated with various neuropsychiatric disorders (e.g., schizophrenia) (41). A recent study investigating DNAm changes in sperm related to human cannabis use reported at least ten differentially methylated CpGs in *DLGAP2* (9). Further, *DLGAP2* exhibits both increased and decreased intronic DNAm, as well as increased mRNA expression in association with parental THC exposure in the rat nucleus accumbens (42).

While prior evidence suggests cannabis-related epigenetic changes in brain for *DLGAP2*, in the case of epigenetic biomarkers for substance use there is no *a priori* expectation that genes dysregulated in blood will reflect neurobiological mechanisms underlying substance use. DNAm profiles differ across tissues and cell types and studies conducted using a more biologically relevant brain tissue (e.g., nucleus accumbens, prefrontal cortex) would be needed to inform the neurobiology of cannabis use disorder (43). Once brain-based DNAm marks in humans (necessitating postmortem tissues) have been identified, however, these results could be used to identify blood proxies of exposure for overlapping blood- and brain-based DNAm changes related to lifetime cannabis use.

Our findings support DNAm as a promising candidate marker for cannabis use. However, our study has limitations. Our findings may not be fully generalizable due to the case-cohort design (i.e., restricted to non-Hispanic Caucasian women and further enriched for women who subsequently developed breast cancer). Although our study relied on self-reported lifetime cannabis use, which could lead to underreporting and

misclassification, these factors would tend to bias towards the null resulting in loss of statistical power. In addition to self-reported lifetime cannabis use, frequency of use was collected; however, the low level of response to this question limited its use to evaluate dose-response. Further, inaccurate and incomplete recall could affect those data, as the average time since first use was 32.6 ± 6.1 years ago with an average total duration of use of 4.6 ± 7.2 years. Time since last use was not collected to assess recency of use. Finally, we cannot rule out the effect of other illicit drugs (e.g., cocaine, opioids) as those data were not collected in the Sister Study. However, we expect the prevalence of other illicit drug use to be low in this cohort. We also do not suspect that our findings are driven by breast cancer, or the use of tobacco and alcohol. In the largest EWAS of tobacco smoking to date (6), cg15973234 was not identified as one of the 18,760 CpGs associated with current and/or former smoking (FDR<0.05), or in the largest EWAS of alcohol to date (44), as one of the 363 CpGs associated with alcohol consumption ($P<1\times10^{-7}$). Further, cg15973234 was not reported as one of the 2,098 replicated breast cancer-associated CpGs in the most recent breast cancer EWAS (45).

Our results provide evidence that blood-based DNAm can inform cannabis use histories and support a surprisingly effective classifier able to categorize ever versus never cannabis users based on DNAm levels at three CpG sites: AUC=0.79 (95% confidence interval [0.76, 0.82]). We anticipate that larger sample sizes and richer phenotype data will enable us to develop more precise DNAm biomarkers, including those indicative of cannabis-related health outcomes.

ACKNOWLEDGEMENTS

This study was supported the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZO1 ES-044005 [to DPS] and ZO1 ES-049033 and ZO1 ES-049032 to JAT), as well as the Fellow Program at RTI International (EOJ). These results have been presented previously at the National Institute of Drug Abuse Genetics Consortium meeting held on Jan 14–15, 2019. We would like to thank Drs. Alexandra White and Kelly Ferguson for their critical review of the manuscript.

DISCLOSURES

Authors have nothing to declare.

REFERENCES

1. (2017): Center for Behavioral Health Statistics and Quality. National Survey on Drug Use and Health: Detailed Tables. *Substance Abuse and Mental Health Services Administration, Rockville, MD*.
2. ProCon.org 33 Legal Medical Marijuana States and DC.: ProCon.org
3. Volkow ND, Compton WM, Weiss SR (2014): Adverse health effects of marijuana use. *The New England journal of medicine*. 371:879.
4. Andersen AM, Dogan MV, Beach SRH, Philibert RA (2015): Current and Future Prospects for Epigenetic Biomarkers of Substance Use Disorders. *Genes*. 6:991-1022.
5. Wilson R, Wahl S, Pfeiffer L, Ward-Caviness CK, Kunze S, Kretschmer A, et al. (2017): The dynamics of smoking-related disturbed methylation: a two time-point study of methylation change in smokers, non-smokers and former smokers. *BMC genomics*. 18:805.
6. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. (2016): Epigenetic Signatures of Cigarette Smoking. *Circulation Cardiovascular genetics*. 9:436-447.
7. Szutorisz H, Hurd YL (2016): Epigenetic Effects of Cannabis Exposure. *Biological psychiatry*. 79:586-594.
8. Gerra MC, Jayanthi S, Manfredini M, Walther D, Schroeder J, Phillips KA, et al. (2018): Gene variants and educational attainment in cannabis use: mediating role of DNA methylation. *Translational psychiatry*. 8:23.
9. Murphy SK, Itchon-Ramos N, Visco Z, Huang Z, Grenier C, Schrott R, et al. (2018): Cannabinoid exposure and altered DNA methylation in rat and human sperm. *Epigenetics*.

10. Sandler DP, Hodgson ME, Deming-Halverson SL, Juras PS, D'Aloisio AA, Suarez LM, et al. (2017): The Sister Study Cohort: Baseline Methods and Participant Characteristics. *Environmental health perspectives*. 125:127003.
11. Kresovich JK, Xu Z, O'Brien KM, Weinberg CR, Sandler DP, Taylor JA (2019): Methylation-based biological age and breast cancer risk. *Journal of the National Cancer Institute*.
12. O'Brien KM, Sandler DP, Xu Z, Kinyamu HK, Taylor JA, Weinberg CR (2018): Vitamin D, DNA methylation, and breast cancer. *Breast cancer research : BCR*. 20:70.
13. Pasman JA, Verweij KJH, Gerring Z, Stringer S, Sanchez-Roige S, Treur JL, et al. (2018): GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia. *Nature neuroscience*. 21:1161-1170.
14. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. (2016): Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology*. 17:208.
15. Xu Z, Niu L, Li L, Taylor JA (2016): ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic acids research*. 44:e20.
16. Xu Z, Langie SA, De Boever P, Taylor JA, Niu L (2017): RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. *BMC genomics*. 18:4.
17. Niu L, Xu Z, Taylor JA (2016): RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. *Bioinformatics*. 32:2659-2663.
18. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. (2012): DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*. 13:86.
19. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. (2013): Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 8:203-209.
20. Venables WN, Ripley BD (2002): *Modern Applied Statistics with S. 4th ed.* New York: Springer.
21. Benjamini Y, Hochberg Y (1995): Controlling for False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 57:289–300.
22. Cohen S, Kamarck T, Mermelstein R (1983): A global measure of perceived stress. *Journal of health and social behavior*. 24:385-396.
23. Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. (2017): Genome-wide methylation data mirror ancestry information. *Epigenetics & chromatin*. 10:1.
24. Tibshirani R (1996): Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 58:267-288.
25. Friedman J, Hastie T, Tibshirani R (2010): Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*. 33:1-22.
26. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. (2011): pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*. 12:77.
27. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. (2017): Disease variants alter transcription factor levels and methylation of their binding sites. *Nature genetics*. 49:131-138.

28. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. (2017): The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*. 26:126-135.
29. DeLong ER, DeLong DM, Clarke-Pearson DL (1988): Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 44:837-845.
30. Yoshida H, Nagaoka A, Kusaka-Kikushima A, Tobiishi M, Kawabata K, Sayo T, et al. (2013): KIAA1199, a deafness gene of unknown function, is a new hyaluronan binding protein involved in hyaluronan depolymerization. *Proceedings of the National Academy of Sciences of the United States of America*. 110:5612-5617.
31. Abe S, Usami S, Nakamura Y (2003): Mutations in the gene encoding KIAA1199 protein, an inner-ear protein expressed in Deiters' cells and the fibrocytes, as the cause of nonsyndromic hearing loss. *Journal of human genetics*. 48:564-570.
32. Marella M, Jadin L, Keller GA, Sugarman BJ, Frost GI, Shepard HM (2018): KIAA1199 expression and hyaluronan degradation colocalize in multiple sclerosis lesions. *Glycobiology*. 28:958-967.
33. Deng F, Lei J, Zhang X, Huang W, Li Y, Wu D (2017): Overexpression of KIAA1199: An independent prognostic marker in nonsmall cell lung cancer. *Journal of cancer research and therapeutics*. 13:664-668.
34. Kohi S, Sato N, Koga A, Matayoshi N, Hirata K (2017): KIAA1199 is induced by inflammation and enhances malignant phenotype in pancreatic cancer. *Oncotarget*. 8:17156-17163.
35. Zhang D, Zhao L, Shen Q, Lv Q, Jin M, Ma H, et al. (2017): Down-regulation of KIAA1199/CEMIP by miR-216a suppresses tumor invasion and metastasis in colorectal cancer. *International journal of cancer*. 140:2298-2309.
36. Petrey AC, de la Motte CA (2014): Hyaluronan, a crucial regulator of inflammation. *Frontiers in immunology*. 5:101.
37. Jiang D, Liang J, Noble PW (2011): Hyaluronan as an immune regulator in human diseases. *Physiological reviews*. 91:221-264.
38. Nagarkatti P, Pandey R, Rieder SA, Hegde VL, Nagarkatti M (2009): Cannabinoids as novel anti-inflammatory drugs. *Future medicinal chemistry*. 1:1333-1349.
39. Liang G, Fang X, Yang Y, Song Y (2018): Silencing of CEMIP suppresses Wnt/beta-catenin/Snail signaling transduction and inhibits EMT program of colorectal cancer cells. *Acta histochemica*. 120:56-63.
40. Boerboom A, Reusch C, Pieltain A, Chariot A, Franzen R (2017): KIAA1199: A novel regulator of MEK/ERK-induced Schwann cell dedifferentiation. *Glia*. 65:1682-1696.
41. Rasmussen AH, Rasmussen HB, Silahtaroglu A (2017): The DLGAP family: neuronal expression, function and role in brain disorders. *Molecular brain*. 10:43.
42. Watson CT, Szutorisz H, Garg P, Martin Q, Landry JA, Sharp AJ, et al. (2015): Genome-Wide DNA Methylation Profiling Reveals Epigenetic Changes in the Rat Nucleus Accumbens Associated With Cross-Generational Effects of Adolescent THC

Exposure. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*. 40:2993-3005.
43. Koob GF, Volkow ND (2016): Neurobiology of addiction: a neurocircuitry analysis. *The lancet Psychiatry*. 3:760-773.
44. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. (2016): A DNA methylation biomarker of alcohol consumption. *Molecular psychiatry*.
45. Xu Z, Sandler DP, Taylor JA (In press): Blood DNA methylation and breast cancer: A prospective case-cohort analysis in the Sister Study *Journal of the National Cancer Institute*.

Table 1. Description of NIEHS Sister Study samples (total N=2,583)[a] by ever/never lifetime cannabis use.

Description	Discovery (N=1,730)			Replication (N=853)		
	Never (N=875)	Ever (N=855)	P-value[b]	Never (N=461)	Ever (N=392)	P-value[b]
Age at blood draw, mean (SD)	60.43 (8.60)	53.63 (7.45)	**2.20×10^{-16}**	60.25 (8.53)	53.15 (7.68)	**2.20×10^{-16}**
Incident breast cancer, N (%)			0.20			0.38
Non-event	369 (42.17)	387 (45.26)		192 (41.65)	175 (44.64)	
Event	506 (57.83)	468 (54.74)		269 (58.35)	217 (55.36)	
Tobacco smoke, N (%)			**6.48×10^{-22}**			**1.86×10^{-9}**
Never	546 (62.40)	336 (39.30)		299 (64.86)	170 (43.37)	
Former	293 (33.49)	432 (50.53)		140 (30.37)	185 (47.19)	
Current	36 (4.11)	87 (10.18)		22 (4.77)	37 (9.44)	
Alcohol use[c], N (%)			**5.65×10^{-11}**			**3.92×10^{-8}**
Noncurrent	197 (22.51)	92 (10.76)		99 (21.48)	31 (7.91)	
Current	678 (77.49)	763 (89.24)		362 (78.52)	361 (92.09)	

[a]Final dataset post-quality control exclusions
[b]P-values are based on a Fisher's exact test and t-test for categorical and continuous variables, respectively. P<0.05 are shown in bold.
[c]Never and former users are combined as there are only N=3 never alcohol users among ever cannabis users

Abbreviation: SD, standard deviation

Table 2. Replication of cg15973234[a]-lifetime cannabis use association.

Sister Study Sample	N	Beta	SE	P-value
Discovery	1,730	-0.009	0.002	1.32×10^{-7}[b]
Replication	853	-0.005	0.003	4.51×10^{-2}
Combined	2,583	-0.008	0.001	3.32×10^{-8}

[a]Probe located within the gene, *CEMIP* (chr15:81072152; GRCh37/hg19 build)

[b]FDR adjusted P-value=0.06

Abbreviation: SE: standard error

Table 3. Independent blood-based *cis*-meQTLs of cg15973234 are not associated with lifetime cannabis use in recent GWAS.

SNP	Gene	Distance to CpG[a]	*cis*-meQTL P-value[b]	GWAS meta-analysis[c]		
				Beta	P-value	MAF
rs3848177	*CEMIP*	0.064	2.37×10^{-18}	-0.0024	0.85	0.16
rs8025670	*ARNT2*	188.6	5.95×10^{-6}	-0.0098	0.41	0.18

[a]Distance in kilobases

[b]N=3,841; BIOS QTL browser (https://genenetwork.nl/biosqtlbrowser/)

[c]Results based on the International Cannabis Consortium (ICC) + UK-Biobank (UKB) samples (N=162,082). These SNPs were not among the top 10K results based on the full sample (N=184,765; ICC + UK Biobank + 23andMe) (Pasman et al. *Nat Neurosci* 2018)

Abbreviation: MAF: minor allele frequency

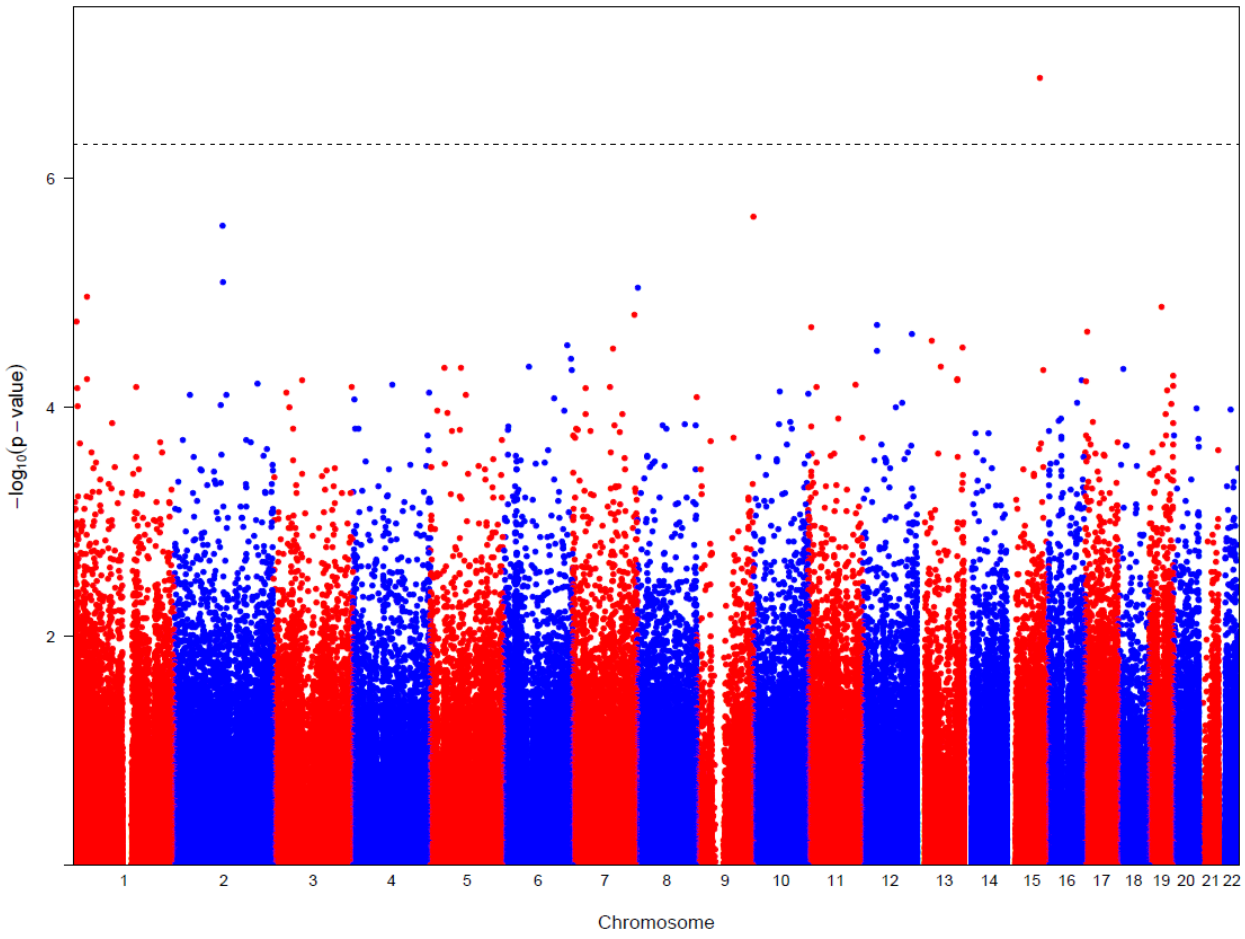

Figure 1. Discovery EWAS of lifetime cannabis use. CpGs are shown according to their position on chromosomes 1–22 (alternating red/blue) and plotted against their -$\log_{10}$P-values. The dotted horizontal line indicates genome-wide significance based on FDR<0.10. The genomic inflation factor (λ) was 1.02.

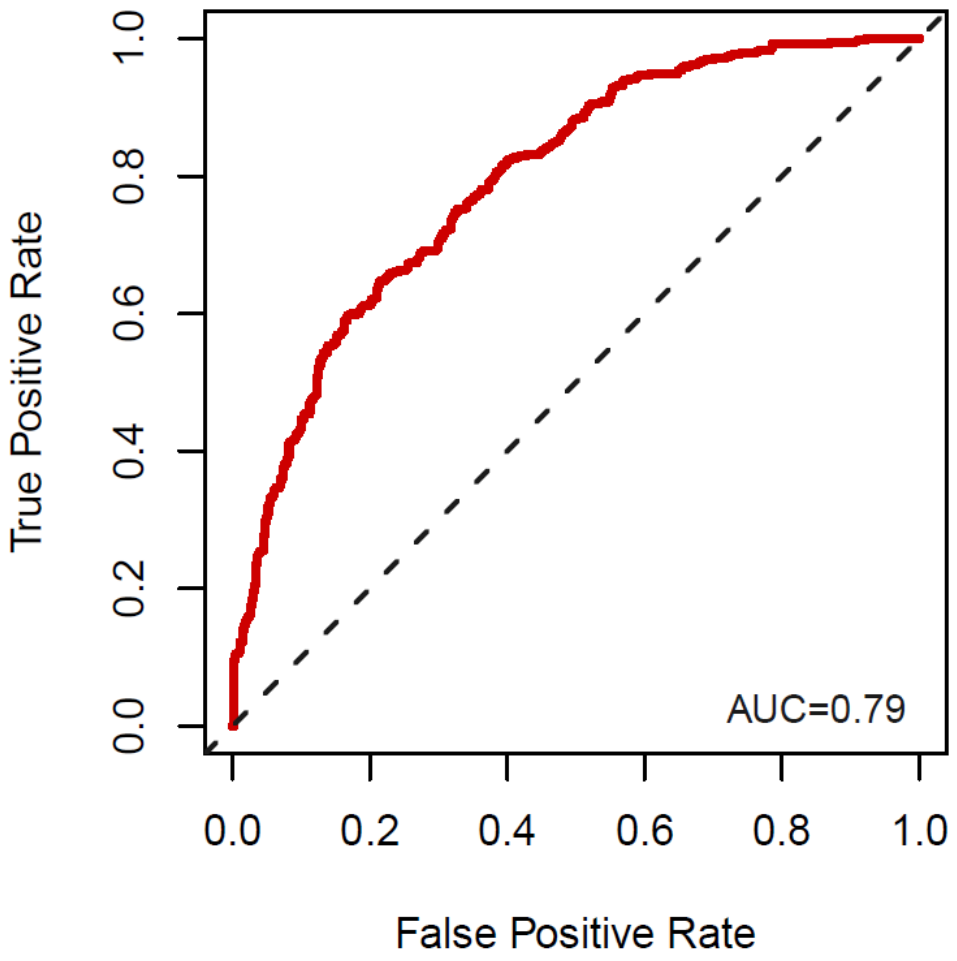

Figure 2. ROC curve of DNAm biomarker (3 CpGs) of lifetime cannabis use in replication sample.