2020

Genetic and environmental prediction of opioid cessation using machine learning, GWAS, and a mouse model

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Dissertation

GENETIC AND ENVIRONMENTAL PREDICTION OF OPIOID CESSATION USING MACHINE LEARNING, GWAS, AND A MOUSE MODEL

by

JIAYI WU COX

B.S., Sichuan University, 2013
M.S., Tufts University, 2015

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader _____

Lindsay A. Farrer, Ph.D.
Distinguished Professor of Genetics
Professor of Medicine
Professor of Neurology
Professor of Ophthalmology
Professor of Epidemiology
Professor of Biostatistics

Second Reader _____

Richard M. Sherva, Ph.D.
Research Assistant Professor of Medicine

The world's first trillionaires are going to come from somebody who masters AI and all its derivatives, and applies it in ways we never thought of. – Mark Cuban

DEDICATION

I would like to dedicate this work to my super loving spouse RJ, my super supporting parents Wenping and Qizhi, and my super encouraging in-laws Susan and John.

ACKNOWLEDGMENTS

The basis for this research originated from my passion of finding a better solution for one to stop using opioid. As more and more digital healthcare data became available for research, there will be a greater need to understand how one's genetics and daily life activities could lead to certain diseases. How will we interpret the data and provide a better personalized management plan to prevent or treat diseases?

In truth, I could not have done such a fun Ph. D project without the support of my PI Dr. Lindsay Farrer, my lab mentor Dr. Richard Sherva, my statistic mentor Dr. Kathryn Lunetta, my addiction clinic mentor Dr. Richard Saitz, my mouse project mentor Dr. Camron Bryant, my machine learning project mentor Dr. Mark Kon, and my dear committee chair Dr. Shoumita Dasgupta. I am also very grateful for my lab mates who helped me significantly in the first few years and taught me various data manipulation techniques. Special thanks to my family and friends for making this journey fun and memorable.

GENETIC AND ENVIRONMENTAL PREDICTION OF OPIOID CESSATION USING MACHINE LEARNING, GWAS, AND A MOUSE MODEL

JIAYI WU COX

Boston University School of Medicine, 2020

Major Professor: Lindsay A. Farrer, Ph.D., Distinguished Professor of Genetics, Professor of Medicine, Professor of Neurology, Professor of Ophthalmology, Professor of Epidemiology, Professor of Biostatistics

ABSTRACT

The United States is currently experiencing an epidemic of opioid use, use disorder, and overdose-related deaths. While studies have identified several loci that are associated with opioid use disorder (OUD) risk, the genetic basis for the ability to discontinue opioid use has not been investigated. Furthermore, very few studies have investigated the non-genetic factors that are predictive of opioid cessation or their predictive ability.

In this thesis, I studied a novel phenotype–opioid cessation, defined as time since last use of illicit opioids (< 6 months ago as not cease, >1 year ago as cease) among persons meeting lifetime DSM-5 criteria for opioid use disorder (OUD).

In chapter two, I identified novel genetic variants and biological pathways that potentially regulate opioid cessation success through a genome wide study, as well as genetic overlap between opioid cessation and other substance cessation traits.

In chapter three, I identified multiple non-genetic risk factors specific to

each racial group that are predictive of opioid cessation from the same individuals analyzed in chapter two by applying several linear and non-linear machine learning techniques to a set of more than 3,000 variables assessed by a structured psychiatric interview. Factors identified from this atheoretical approach can be grouped into opioid use activities, other drug use, health conditions, and demographics, while the predictive accuracy as high as nearly 80% was achieved. The findings from this research generated more hypotheses for future studies to reference.

In chapter four, I performed differential gene expression and network analysis on mice with different oxycodone (an opioid receptor agonist)-induced behaviors, and compared the significantly associated genes and network modules with top-ranked genes identified in humans. The pathway cross talks and gene homologs identified from both species illuminate potential molecular mechanism of opioid behaviors.

In summary, this thesis utilized statistical genetics, machine learning, and a computational biology framework to address factors that are associative with opioid cessation in human, and cross referenced the genetic findings in a mouse model. These findings serve as references for future studies and provide a framework for personalizing treatment of OUD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAs .....African Americans

ATP..... adenosine triphosphate

cAMP .....cyclic adenosine monophosphate

CATS ..... Comorbidity and Trauma Study

CDC..... U. S Centers for Disease Control and Prevention

CPP .....conditioned placed preference

DNN ..... deep neural network

DSM-5 .....Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

EAs ..... European Americans

EHR .....electronic health record

EPM ..... elevated plus maze

GA .....genetic algorithm

GEE ..... generalized estimating equation

GPCR ..... G-protein coupled receptor

GSEA ..... gene set enrichment analysis

GTEx .....Genotype Tissue Expression Portal

GWAS ..... Genome-wide association studies

IBD .....identity-by-decent

ICD-10-CM

..... International Classification of Diseases, 10th Revision, Clinical Modification

LASSO .....least absolute shrinkage and selection operator

LD .....linkage disequilibrium

MAF .....minor allele frequency

NAS ..... neonatal abstinence syndrome

OD ..... opioid dependence

OST ..... opioid substitution therapy

OUD.....opioid use disorder

PC ..... principal component

PRS ..... polygenic risk score

QTL ..... Quantitative trait locus

RF .....random forest

RFE .....recursive feature elimination

SNP .....single nucleotide polymorphism

SSADDA.....Semi-Structured Assessment for Drug Dependence and Alcoholism

SUD .....substance use disorder

SVM .....support vector machine

UKBB ..... UK Biobank

CHAPTER ONE: Introduction

Current epidemic of opioid use disorder (OUD[1])

Socioeconomic impact

Medical opioids were prescribed at an increasing rate in the mid to late 1990s, which was followed by a dramatic rise of potent, synthetic, non-medical opioid use[2]. Opioid overdose is now the leading cause of accidental death in the United States (U.S.)[3]. More than half a million people in the U.S. died from opioid overdose from 2000-2015[3]. In addition to the rapid increase of overdose mortality, illicit opioid use also resulted in growing rates of emergency room visits, babies born with symptoms of drug withdrawal – termed neonatal abstinence syndrome (NAS), and Opioid Use Disorder (OUD) treatment admissions[4]. In 2011, the U. S Centers for Disease Control and Prevention (CDC) officially declared an opioid epidemic, with the Northeast, Midwest, Appalachia, and many southern states being the most severely affected regions (Figure 1.1)[5]. In 2016, more than 90 Americans died from opioid overdoses every day, an increase of 21% compared to 2015[2]. From 2010 to 2014, heroin involved overdose death has tripled[6].

The total economic cost of the opioid epidemic is stunning. The Council of Economic Advisers (CEA) of the Office of the President of the United States published their analysis on the economic burden created by the opioid crisis in 2015. The total cost of the opioid epidemic was estimated at $504 billion, considering both prescription and non-prescription opioid use, which includes

$72.3 billion for non-fatal consequences (healthcare cost for OUD therapy, criminal-justice cost, reduced productivity) and $431.7 billion in fatality costs[7]. The striking cost estimated for the opioid epidemic highlights for both policy makers and researchers the stark need for a better understanding of various interventions to mitigate this crisis.

Persons with low income, low education, and limited employment opportunities are more likely to be the victims of opioid addiction. In 2012, 81% of hospital charges for NAS were covered by Medicaid programs, a federal and state program that helps with medical cost for people with limited income[8]. While opioid overdose death rates have increased for all education groups, those with less than a high school education lost the most years from their estimated life expectancy. Persons with lower education might have fewer incentives to quit and have more risk factors for engaging in drug use than their more educated counterparts[9]. An association between opioid use and unemployment has also been observed, where the opioid crisis was thought to be more severe in areas with high rates of unemployment[10].

Opioid classification and mechanisms of action

There are several ways to classify opioids. Categorizing by synthetic process, there are as naturally occurring compounds (such as morphine and codeine), semi-synthetic compounds (such as diamorphine – also known as heroin, buprenorphine, and oxycodone), or synthetic compounds (such as fentanyl and methadone)[11]. Opioids can also be classified by their effect on

opioid receptors. Opioid agonists (e.g., morphine, diamorphine, codeine, fentanyl, methadone, oxycodone) bind tightly to opioid receptors and cause a significant conformational change to produces a maximal response, whereas partial agonists such as buprenorphine activate opioid receptors to a much smaller degree than agonists. Opioids can also be categorized based on the types of receptors they bind. All opioid receptors are G-protein coupled receptors (GPCR) which include mu, delta and kappa opioid receptors[12]. Most opioids used in clinical practice act mainly through mu opioid receptors, which are responsible for generating opioids' analgesic effects.

On the molecular level, opioid receptor activation leads to a series of downstream cascades, including inhibition of neurotransmitter release, reduction of calcium influx, escalation of potassium efflux, and inhibition of adenylate cyclase, an enzyme that breaks down adenosine triphosphate (ATP) to form cyclic adenosine monophosphate (cAMP)[13]. When an opioid binds to its GPCR, the alpha subunit of GPCR disassociates from the beta and gamma G-protein complex. The disassociated alpha subunit interacts with the G-protein gated potassium channels and leads to increased potassium efflux, while the beta and gamma G-protein complex inhibits calcium conductance by directly binding to the channel, causing cellular hyperpolarization resulting in neural activity inhibition. The C-terminal domain of the GPCR is further phosphorylated upon opioid binding, which recruits and binds to the adapter protein beta-arrestin[14]. The phosphorylated arrestin-bound GPCR activates MAPK pathways, which

participate in multiple cellular processes, including cell proliferation, differentiation, apoptosis, transcription factor regulation, and protein scaffolding[15].

The positive reinforcement (reward) of opioid results primarily through the dopamine system. At the physiological level, dopaminergic neurons are located mainly in the mid brain ventral tegmental area. Their axons project to the striatum (including nucleus accumbens and putamen), amygdala, and prefrontal cortex[15]. The ventral tegmental area and nucleus accumbens represent the mesolimbic dopaminergic reward system, which mediates opiate reinforcement[16]. Opioids when bound to mu receptor, it causes a disturbance of potassium channel activity and results in a reduction of excitatory neurotransmitters such as substance P and glutamate release in and from the prefrontal cortex[15,17]. The neuron hyperpolarization and reduced release of the inhibitory neurotransmitters contribute to opioid's analgesic effect and addictive properties.

Pharmacological actions of opioid agonists have been well studied. Opioids have a strong effect on the central nervous system and result in analgesia, sedation, euphoria, dysphoria, tolerance and dependence, and also impact the cardiovascular system leading to mild bradycardia and peripheral vasodilatation. They also affect the respiratory system, leading to respiratory depression, cough suppression; the gastrointestinal system, leading to nausea and vomiting; the muscular system, leading to muscle rigidity and pupil constriction; and the immune system, leading to immune depression[18].

Clinical criteria for OUD and current treatment

The 2018 International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM)[19] and American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) both have a set of complex criteria to quantify opioid use and its related conditions. DSM-5 considers 11 OUD symptom criteria that account for loss of control, social problems, risky use, and pharmacological problems. An OUD diagnosis requires that a person meet at least two criteria within a 12-month period. Severity of OUD is defined by the number of symptoms as follows: mild (2-3 symptoms), moderate (4-5 symptoms), or severe (>6 symptoms) (Table 1.1)[20]. According to DSM-5, early opioid remission lasts 3-12 months and sustained remission requires a period of at least 12 months without any OUD diagnostic criteria (except for craving)[21]. ICD-10-CM opioid related disorder criteria (Table 1.2)[22] are generally similar to DSM-5, for example, there are two remission codes corresponding to opioid use severity. However, ICD-10-CM does not define withdrawal for opioid abuse[23].

There are various behavioral and pharmacological interventions to treat OUD. Behavioral interventions include help by well-trained counselors to identify risk behaviors[24] and by patient advocates, and self-help[25]. Pharmacological interventions mostly target mu opioid receptors. Methadone, for example, is a long-acting opioid agonist that has been shown to reduce relapse rate, facilitate behavioral therapy, and enable OUD sufferers to live a normal life[26]. Medical use

of methadone produces minimal tolerance, reduces craving, and has been associated with positive long term outcomes[27,28]. Naltrexone is a mu receptor antagonist; it lessens the potential for relapse by tightly binding to the mu opioid receptor without producing pleasure. Naltrexone has lower compliance than methadone because one must be fully withdrawn from all opioids agonists to avoid withdrawal[29]. Buprenorphine is a mu receptor partial agonist; patients with opioid physical dependence (display symptoms such as gastrointestinal upset or gooseflesh skin) are likely to withdrawal.

Figure 1.1 Drug overdose death rate by state[5].

Northeast, Midwest, Appalachia, and several southern states are the most severe regions.

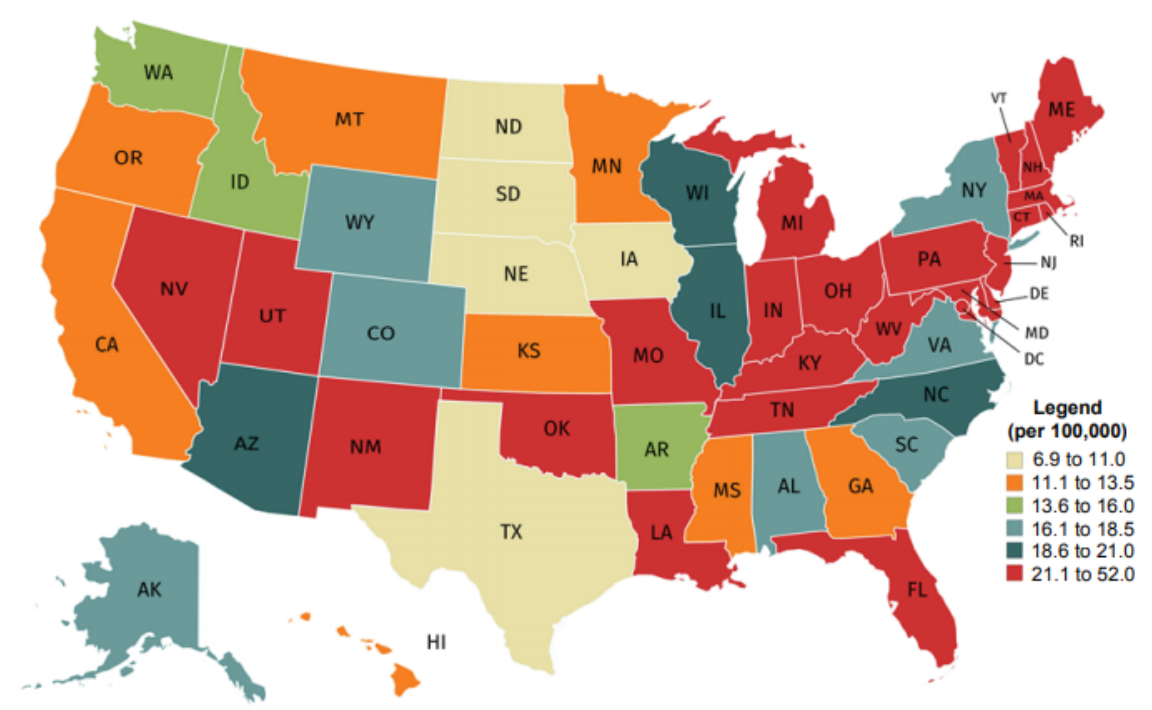


Table 1.1 DSM-5 OUD criteria[20].

OUD is defined as a problematic pattern of opioid use that leads to serious impairment or distress. OUD patients must meet at least two of the following symptoms within a 12-month period. Mild OUD: 2-3 symptoms, moderate OUD: 4-5 symptoms, severe OUD: more than 6 symptoms.

Loss of Control	
1	Substance taken in larger amounts or for a longer time than intended
2	Persistent desire or unsuccessful effort to cut down or control use of a substance
3	Great deal of time spent obtaining, using, or recovering from substance use
4	Craving (a strong desire or urge) to use opioids
Social Problems	
5	Continued opioid use that causes failures to fulfill major obligations at work, school, or home
6	Continued opioid use despite causing recurrent social or personal problems
7	Important social, occupational, or recreational activities are reduced because of opioid use
Risky Use	
8	Recurrent opioid use in dangerous situations
9	Continued opioid use despite related physical or psychological problems
Pharmacological Problems	
10	Tolerance (the need to take higher doses of a drug to feel the same effects, or a reduced effect from the same amount)
11	Withdrawal (the experience of pain or other uncomfortable symptoms in the absence of a drug)

Table 1.2 ICD10 opioid dependence criteria[22].

1	a strong desire or sense of compulsion to take opioids
2	difficulties in controlling opioid-use behaviors in terms of the onset, termination or levels of use
3	a physiological withdrawal state when opioid use has ceased or been reduced, as evidenced by one of the following
	a. the characteristic withdrawal syndrome
	b. use of opioids (or closely related substances) with the intention of relieving or avoiding withdrawal symptoms
4	evidence of tolerance, such that increased doses of opioids are required to achieve effects originally produced by lower dose
5	progressive neglect of alternative pleasures or interests because of opioid use; increased amounts of time spent to obtain opioids or to recover from their effects
6	persisting with opioid use despite clear evidence of overtly harmful consequences, such as depressive mood states consequent to periods of heavy substance use, or drug-related impairment of cognitive functioning (efforts should be made to determine that the user was actually, or could be expected to be, aware of the nature and extent of the harm)

Studies of genetic and phenotypic factors relating to OUD in human

Studies of human genetics overview

Linkage and candidate gene studies

To identify genes that are associated with addictive traits, a good first step is to evaluate evidence of heritability of such traits. More than two decades ago, genetic linkage analysis and candidate gene studies were popular tools used to identify chromosome regions that are associated with a trait. Linkage studies make it possible to identify chromosomal locations of the disease-associated genes based on the fact that the shorter the physical distance between two genes on a chromosome, the less likely recombination will occur between them during meiosis. Linkage studies are particularly useful in diseases that are inherited in a Mendelian fashion, but such studies require data from multiple family members including at least two affected persons and, thus, are often limited by sample size[30]. Regions on chromosome 17 in African Americans (AAs), European Americans (EAs), and Han Chinese have been linked to opioid dependence[31,32].

Linkage studies can establish the genetic basis of a trait, but often lack the resolution necessary to point to specific genes involved in a trait. In contrast, hypothesis-based candidate gene studies focus on genes that have been previously related to the disease experimentally. By comparing the frequency of single nucleotide polymorphisms (SNPs) in persons with disease and controls, one can evaluate whether a gene is associated with the disease[33]. However,

several limitations should be noted such as the possibility of false positive associations due to population admixture, multiple testing correction issues due to testing the same SNP in different traits or multiple SNPs for the same trait, and selection bias due to limited and sometimes incorrect knowledge about physiology[30]. The mu-opioid receptor (*OPRM1*), the delta-opioid receptor (*OPRD1*), the dopamine D2 receptor (*DRD2*), and brain-derived neurotrophic factor (*BDNF*) have been extensively studied as candidate genes for opioid dependence in several populations[34-36], the polymorphism of these genes have shown to be affecting the traits.

Genome-wide association studies (GWAS)

With the development of microarray chip technology, an extremely large number of markers can be assessed for association with disease in a single experiment. GWAS use hypothesis-free methods for identifying associations between genetic loci and traits. GWAS usually evaluate association of disease status with common SNPs and rely on linkage disequilibrium (LD), population level links between the SNPs and the causal markers, which may or may not be tested directly[37]. Given the millions of markers tested directly or imputed using LD information from population reference panels, an alpha level of 5×10^{-8} is generally used as a significance threshold to correct for the large number of tests. GWAS require thousands or tens of thousands of samples to achieve adequate power to identify SNPs with a small effect on disease risk, and the true causal variants related to the association signal might not be located in the genes

containing the top-ranked variants. The number of signals identified by GWAS for a specific trait depends on the number of loci affecting the trait, the allele frequencies and effect sizes of the loci, sample size, and trait heterogeneity[38]. Potassium voltage-gated channel subfamily G member 1 (*KCNG2)* was the first gene identified from GWAS for opioid dependence in both AAs and EAs[39]. Subsequently, a SNP near repulsive guidance molecule BMP co-receptor A (*RGMA*) was associated with opioid dependence in an EA sample[40].

Rare variant studies

Although numerous discoveries have been made identifying common variants of OUD through GWAS, much of the heritability remains unexplained. Many Mendelian disorders and rare diseases are caused by highly penetrant rare variants[41]. Natural selection tends to eliminate deleterious alleles, and loss of function variants are especially rare[42]. Next generation sequencing and high confidence genotype imputation have made it possible to analyze infrequent variants. Given the limited statistical power of a single variant analysis approach using GWAS, several statistical methods were developed for rare variant analysis. These methods include (1) the burden test, which aggregates rare variants into gene- or region-based scores that are used for an association test; (2) the adaptive burden test, which adds adaptive weights on top of the burden test to magnify the effect of deleterious variants; (3) variance-component test, which tests the variance of genetic effects; and (4) the combined test, which combines the benefit of both burden and variance-component tests[43]. However,

to date there are not any publications studying rare variants in opioid related traits.

Enriched pathways analysis

Associations with variants and genes obtained from GWAS usually do not provide a sufficient explanation of disease mechanisms when considering genes individually. Other methods can translate these findings into a more biologically coherent understanding of the underlying disease processes. Enriched pathway analyses, including gene over-representation and aggregate gene scores, can help with this. Both analyses form gene sets based on a biological process or certain molecular functions derived from information in existing databases such as Gene Ontology[44]. Steps to carry out the over-representation analysis include forming a list of top genes from GWAS, comparing the number of GWAS genes that are members of each gene set to those who are not members using hypergeometric distribution or binomial approximation to infer statistical significance[45]. Many of these methods have been implemented as web-based tools, such as Enrichr[46]. As a limitation, the over-representation method ignores genes that are not in the GWAS selected gene list and the result depends on the cutoff for the gene selection. The aggregate score approach overcomes this limitation by including all the genes from GWAS in the analysis. One of the most representative methods is gene set enrichment analysis (GSEA), which aims to calculate if gene scores in gene set S are different from gene scores in genes outside of S. For GSEA, each SNP is assigned to a gene followed by

computation of SNP-based gene scores, then it compares the rank of the genes included in a pre-specified gene set to the rank not included in the gene set, based on the Kolmogorov Smirnov test[45]. There are several limitations of the enriched pathway analysis approach. For example, the use of a gene set might be limited by current biological knowledge; furthermore, many GWAS SNPs may not be located in genes or may overlap several genes. Using enriched pathway analysis, genes identified through opioid dependence case control studies are overrepresented in calcium and potassium pathways [39].

Human SUDs behavioral research through large epidemiology studies

Various epidemiology studies use survey data to derive DSM-5 criteria to classify types of SUDs in susceptible groups, such as adults[47,48],people with comorbid psychiatric diseases[49], people from different geographical regions[50], and unique social groups[51,52]. Although the goals of these large epidemiological studies are to identify factors that influence the distribution and etiology of SUDs for prevention or intervention efforts, a limited amount of factors (e.g. sociodemographic characteristics, comorbidity, disability, and treatments) were previously analyzed[48]. Moreover, many of these studies use simple statistical methods such as regression or correlation to determine the influence of a specific factors [48,52]. With the increasing availability of electronic health records (EHRs) for epidemiologic research, more opportunities are available for researchers to access high dimensional data and identify additional factors associated with

SUDs. Such datasets with thousands of variables might not be efficiently analyzed with simple regressions and would suffer from overfitting and insufficient power issues due to the increasing number of covariates[53].

Deriving patterns from the high dimensional datasets using machine learning

Data mining using machine learning is increasingly used to identify abnormalities, classify patients, facilitate diagnoses, and evaluate prognoses. Feature selection as a data mining method can be effectively used to remove irrelevant data, increase learning accuracy, and improve comprehensibility.

Since an exhaustive search for optimal feature subsets is not feasible in most cases, alternative strategies have been proposed in three main categories: filter, wrapper, and embedded methods[54]. The filter methods select features based on a performance measure independent of the prediction/classification method used in the later stage. Information gain, Euclidean distance, and p values from the univariate tests are examples of such performance measures. Wrapper methods consider feature subsets by the quality of performance on a modeling algorithm. The evaluation is performed on each subset using the same search rule, such as recursive feature elimination (RFE) [55]. Wrapper methods have been proven to perform better than filter methods while taking longer time to execute. Embedded methods are greedy search based algorithms that perform intrinsic feature selection. One of the popular techniques in embedded

methods is stepwise regression. Depending on the model, one might start with an empty list and recursively add variables one by one or start with a full list of variables and recursively remove variables one by one until the variable set collectively achieves the best fit[56]. Limitations of using stepwise regression include its limited ability to find the global optimization, method instability with different models, and the maximum number of variables that can be considered [56,57].

Regression based models have been used for estimating the relationship between a single dependent variable with one or more independent variables. However, there are limitations to regression based methods such as the assumption of linearity between the independent and dependent variables, the assumption of little or no multicollinearity between dependent variables, and normally distributed residuals. With the emergence of other machine learning algorithms such as neural networks, support vector machines, and random forests, several of these limitations have been addressed. The logic behind some of the popular feature selection methods employed in Chapter three are described in Table 1.2.

Logistic Least absolute shrinkage and selection operator (LASSO)

LASSO is a regression based model that uses L1-norm (absolute value) regularization to shrink the effect estimate of irrelevant variables to zero and it has been popularly used in biological studies [58]. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficient being

less than a constant[59]. LASSO is widely used especially for problems where the number of variables p exceeds the number of observations n. However, when $\mathrm{p} > n$, the LASSO criteria is not strictly convex and thus might yield an unstable solution[60]. During feature selection, LASSO randomly picks one variable from a group of correlated variables.

Linear–Support vector machine (linear-SVM)

SVM is a discriminative classifier best known for the 'kernel trick' – the ability to transform the value of each feature into a different dimension via a kernel function -- and is widely used in high dimensional genomic data[61,62]. SVM finds the best linear separating hyperplane with the maximal margin between two classes of data. RFE has proven to yield better classification performance using SVM [63]. To avoid overfitting, a regularization parameter term is added to allow for a certain degree of misclassification.

Random forest (RF)

RF is an ensemble classification and regression method that yields high classification accuracy. RF is able to achieve both low bias and low variance by averaging across multiple low bias, high variance decision trees. Each decision tree is constructed by sampling with replacement from the available training data pool. Gini impurity, a measurement based on information criteria, is used to indicate the feature importance[64]. Variables that lead to higher reduction in Gini impurity are ranked higher than those who don't. Similar to linear-SVM, RFE is

used to select features based on the Gini importance of the variables.

Deep Neural Network (DNN)

DNN uses multilayer perceptron to process knowledge and make predictions. The structure and function of DNN are based on our understanding of the nervous system, where neurons are interconnected in such a way that they can store information and determine if a signal is strong enough to be passed on. A standard DNN contains an input layer, where each node takes in a set of information from the input; one or several user-defined hidden layers, where every node from each layer is fully connected to all the nodes from the previous layer, as well as to all the nodes from the next layer; and an output layer that yields the prediction result. As the building block of DNN, each node performs the sum of a linear combination from the inputs of the previous layer with certain weights and biases. After summation, information is subject to a linear or non-linear transformation by an activation function. This step allows DNN to model any linear or non-linear relationships between the inputs and outcome. Feature selection in DNN has been studied by several groups, and more details will be described in chapter three.

Other methods

Other feature selection methods using machine learning, such as genetic algorithms (GAs) and hybrid methods, should also be noted. GA is a stochastic method for finding optimization using the theory of natural evolution and is best

known for its ability to efficiently search large spaces with little known a priori[65]. A user-defined classifier (DNN, SVM, etc.[66]) can be used for GA. An initial set of variables are randomly picked by each classifier and trained on a training set. The performance (fitness) of each classifier is then evaluated using an independent dataset. Based on the fitness of each classifier, top classifier-associated features are picked for the next round of evaluation while new sets of features are created by “crossing over” or “mutating” features from these classifiers. “Crossover” and “mutation” are special strategies used by GA in order to find the optimized feature space; both strategies adopt partial inputs from the high-performing classifiers and generate more feature combinations for further testing. The fitness evaluation, selection, crossover, and mutation steps are repeated until a stopping criterion is met, where the inputs from the most fitted classifier would contain the most relevant features. GA does not require any a priori knowledge of the problem under investigation, but is very computationally expensive. We did not pursue this method given that GA iteratively builds many predictive models and can take a long time to converge. The hybrid method is an application of a combination of several machine learning methods, such as linear-SVM coupled with LASSO, DNN coupled with linear-SVM, etc. The hybrid method application is left as a future direction to explore possible improvements based on what was observed using the primary methods.

Table 1.3 Characteristics of the selected machine learning methods.

Pros and cons of LASSO, linear-SVM, RF and DNN as feature selection methods. LASSO is an embedded method that performs feature selection and outcome prediction simultaneously. Both linear-SVM and RF uses a wrapper method – recursive feature elimination to select important variables. For DNN, we used a filter based method, which picks the inputs based on their net contributions to the next node (see Chapter three). LASSO: Logistic Least absolute shrinkage and selection operator, linear-SVM: linear support vector machine, RF: random forest, DNN: deep neural network.

	Pros	Cons	note
LASSO	sparsity, leaves few correlated variables, parametric, model inter-variable relationships, simpler parameter tuning, provides directionality between the predictors and the outcome	unstable, more biased than OLS	linear
Linear-SVM	uses "support vectors" to make decision, has guaranteed optimality, provides directionality between the predictors and the outcome	Nonparametric, picks correlated variables, non-parametric, less effective in noisy data	linear
RF	An ensemble method, uses how often a feature is selected for splitting 2 classes as a criteria for feature importance	Nonparametric, picks correlated variables, no directionality between the predictors and the outcome	non-linear
DNN	models intricate inter-variable relationship (e.g interactions), powerful in detecting complex patterns, framework allows flexible sub method selection	Nonparametric, picks correlated variables, low interpretability, trial and error-based, more hyperparameters to tune, not easy to find the global minimum, requires a higher volume of data, no directionality between the predictors and the outcome	non-linear

Studies of genetic and phenotypic factors relating to OUD in non-human model organisms

Overview of non-human model organism studies

Animal model organisms for addiction studies

Experimenters are able to use animals to model distinct components of the addiction process, and evaluate a range of behaviors from simple (e.g. acute drug responses) to more complex behaviors (e.g. drug seeking, self-administration, and relapse). Genetic studies in model organisms (such *as C. elegans, Drosophila*, zebrafish, and mouse) have led to the identification of novel and previously unsuspected genes affecting addictive processes. *C. elegans* has many attributes that make it a useful model for studying neurobiological aspects of addiction, for instance it has a remarkably small and well-characterized nervous system, fully sequenced genes, short reproduction cycle (3 days), and the ability to self-fertilize[67]. Similar to *C. elegans*, *Drosophila* models are low cost and size, and relatively easy to handle[68]. Around 75% of human disease genes have homologs in *Drosophila*[68]. Behavioral assays in zebrafish allow exploration of acute drug response and drug associated reinforcements, such as drug-induced locomotor activity and withdrawal behaviors[69].

Mice as a commonly used model organism for addiction

Mammalian model organisms offer the most comprehensive set of genomic and molecular tools for identifying addiction related traits [70,71]. 99% of

mouse genes have human homologues while 95% of mice variants have human counterparts[72], which allows for translation of gene function to humans[73]. Mice are relatively easy to breed and require a limited amount of space to house[73,74]. Importantly, environmental factors and drug administration can be precisely controlled, allowing more accurate research on certain relevant genetic factors. Mice behavioral paradigms are well-developed, some of which permit rapid collection of phenotypes on a relatively large scale. Mice and humans show a considerable amount of overlap in neural pathways associated with addictive drugs[75,76]. Gene editing tools such as CRISPR allow one to rapidly develop models with a desired genotype for demonstrating genetic causality[77,78].

However, it is still challenging to model the complexity of addiction in humans using mouse models. While we can model aspects of the human condition of drug seeking and taking, for example, by manipulating economic cost (e.g. demand curve analysis) or drug availability (schedules of access), these do not perfectly mimic many factors of addiction that are unique to human society (e.g. continued use despite social, legal, or financial consequences).

There is a large amount of methodological overlap between mice and humans when it comes to studying the genetic underpinning of addiction traits. On top of the genetic approaches listed above, described below are additional genetic methods that are commonly used in mice to study addiction traits.

Quantitative trait locus (QTL)

QTL mapping locates genetic loci and estimates their effect in regulating quantitative traits. The idea of QTL mapping is similar to linkage studies in humans, where the genetic variants responsible for addiction are detected by the statistical association between the variations and genetic markers in backcrossed or intercrossed progeny derived from crossing two inbred strains[79]. The discovery potential of QTL mapping is limited to the polymorphism sites between the inbred parental strains, and as a result, the QTL intervals are often too large to identify causal variants for a lower generation of mouse strains (such as F2 cross) because the strains have only experienced a limited amount of recombination events. Several QTL studies have studied loci associated to opioid phenotypes using inbred mouse strains[80,81], specifically, one study has revealed an interval on chromosome 11 associated with opioid sensitivity[82], and several candidate genes on chromosome 7 relating to opioid response[83].

Transcriptomic and gene network analysis

Using animal models allows use of RNA-sequencing techniques to investigate gene expression profiles in multiple tissue types at multiple time points in response to drug treatment or in subjects with different genetic backgrounds[84]. The recent development of single-cell RNA-sequencing allows for the profiling of gene expression patterns in individuals cells and the classification of different neural cell types[85]. Transcriptomic analysis also includes the study of alternative splicing, where the different genotypes of interest or the drug

treatment could lead to increased proteome diversity and regulate protein function[86]. Opioid dependent transcriptional response has been widely studied in several neural cell types in mice[87,88].

Gene network analysis is an additional way to yield biological insight from a list of pre-selected genes. Constructing networks from co-expressed genes can lead to new discoveries in gene functions, gene-gene interactions, and potential biological mechanisms[89]. For example, immune-signaling and ERK1/2 were found as novel genetic markers for multiple addiction phenotypes through gene network analysis using genes previously identified[90].

Mice behavioral studies

There are many methods available for studying the incentive, learning, and drug reward behaviors in rodent models, which can further our understanding of the neural basis of addiction. Below are a subset of behavioral assays in mouse models that can be used to investigate drug sensation, rewarding effect, and withdrawal behaviors that may translate to humans.

Rewarding and aversive effects of drugs

Conditioned placed preference (CPP) is a preclinical behavioral test to study drug reward and aversion. Animals are trained to associate distinct environments with drug/non-drug treatment. On the test day, under drug-free conditions, animals are given the opportunity to choose which environment they prefer to spend the most time, adjusting for the baseline preference. The CPP

paradigm requires little training when compared to self-administration procedures and can be combined with other techniques to elucidate the subjective effects of drugs[53].

Drug induced locomotor activity

Locomotor activity is used to assess whether a drug possesses psychostimulant or sedative effects by measuring the distance the experimental animal traveled in the open field after administration of the drug. However, while opioids induce sedation in humans[91], they can have either sedative or stimulant effects in mice or rats based on dosage[92,93]. Although drug induced locomotor activity is sometimes used as a proxy to evaluate the drug's rewarding effect, locomotor activity and rewarding effects of drugs are regulated by different circuitries in mice[94].

Drug withdrawal behaviors

Opioid withdrawal is known to induce anxiety in humans. Elevated plus maze (EPM) is a test used to evaluate anxiety-like behavior induced by drug withdrawal. EPM consists of two open arms and two closed arms of equal length that are oriented along a single plane and elevated a certain distance above the floor[95]. Previous research has suggested that both spontaneous and naloxone-precipitated withdrawal from opioids in rats causes anxiogenic-like effects (reduction time spent in the open arms) in the EPM[96]. Mice drug withdrawal induced anxiety-like behaviors are displayed differently, where they tend to

spend longer time in the open arms during both spontaneous and naloxone-precipitated opioid withdrawal[97,98].

Specific aims

The primary goal of this thesis project is to identify genetic and phenotypic factors underlying opioid cessation, specifically the ability of people who meet DSM-5 OUD criteria to quit opioid use for at least one year. The three aims of this project (listed below) combine the use of GWAS, machine learning methods, and a mouse model with chronic opioid treatment as research tools to find potential variants, genes, pathways, networks, and behavioral variation that provide insights into understanding opioid cessation.

Aim 1: Identify common variants associated with opioid cessation through GWAS performed using multi-ethnic (AAs and EAs) samples, and strengthen the most promising findings in an independent EA sample. Potential biological pathways influencing opioid cessation will be evaluated using enriched pathway analysis that incorporates the GWAS summary information. In addition, the genetic overlap between opioid cessation and other SUD traits using GWAS results obtained from a large population-based sample deposited in the UK Biobank.

Aim 2: Identify phenotypic factors that are predictive of opioid cessation by applying several machine learning methods to analyze several thousand of variables related to drug use, mental and other medical conditions, and demographic and lifestyle factors obtained from 2557 EA and 1192 AA

individuals meeting DSM-5 criteria for OUD.

Aim 3: We used a mouse model to investigate possible opioid cessation genes; particularly, we employed a reduced genetic complexity cross between mouse substrains that display differential opioid sensitivity and withdrawal behaviors. Given the limitations in modeling human opioid cessation using mice, we aimed to cross- referenced trans-regulated genes and networks associated with a QTL regulating opioid sensitivity to human opioid cessation genes that were identified from GWAS. This framework provides an indirect way to potentially validate genes identified from association studies in humans by generating hypotheses regarding common biological pathways and can be tested *in vivo*.

CHAPTER TWO: GWAS of Opioid Cessation

Abstract

The United States is currently experiencing an epidemic of opioid use, use disorder, and overdose-related deaths. While studies have identified several loci that are associated with OUD risk, the genetic basis for the ability to discontinue opioid use has not been investigated. We performed a genome-wide association study (GWAS) of opioid cessation in 1,130 African Americans (AAs) and 2,919 European ancestry (EAs) participants recruited for genetic studies of opioid, cocaine, or alcohol use disorders and who met DSM-5 criteria for OUD. Opioid cessation was defined as abstinence from opioids for at least one year before the interview date. We examined the association of opioid cessation status genome-wide using generalized estimating equations to account for correlations among related individuals and a model that included terms for imputed single nucleotide polymorphism (SNP) dosage, sex, age and the first five principal components of ancestry. Association tests performed separately within each ethnic group were combined by meta-analysis with summarized results obtained from the Comorbidity and Trauma Study. Although there were no genome-wide significant associations, we found suggestive associations with nine independent loci, three of which were of biological relevance: rs4740988 in *PTPRD* ($p=2.24\times10^{-6}$ in AAs and EAs combined), rs36098404 located in the intron of *MYOM2* ($p=2.24\times10^{-6}$ in EAs), and an intron variant rs592026 of *SNAP25-AS1* ($p=6.53\times10^{-6}$ in EAs). Pathway analysis identified significant pathways in EAs that are related to vitamin

D metabolism ($p = 3.79\times10^{-2}$) and FGF signaling ($p = 2.39\times10^{-2}$). We also found evidence of shared genetic underpinnings between opioid cessation and risk of other substance use disorders, cessation of smoking and drinking, and chronic back pain in the UK Biobank by evaluating the association of polygenic risk scores constructed from summary statistics of these phenotypes with opioid cessation. These results provide evidence for genetic influences on opioid cessation, suggest genetic overlap with other relevant traits, and may indicate potential novel therapeutic targets for OUD.

Introduction

Non-prescribed use of opioid analgesics has become a significant global problem that affects the health and economic welfare of society[99]. In 2017, more than 47,000 Americans died of an opioid overdose, and about 40% of these deaths involved prescription opioids[100]. The U.S Department of Health and Human Services declared a public health emergency in 2017 to address the national opioid crisis[101]. Opioid maintenance and cognitive behavioral approaches are effective[102,103], but they have limited long-term value reflected by high dropout and relapse rates[104]. The nearly 80% of Americans with opioid use disorder (OUD) who do not seek treatment[6] adds to the burden of this devastating public health problem. Several studies have identified genetic variants related to OUD[39,105], which is highly heritable ($h^2 = 0.43,0.50$)[106]. The lack of published reports about genetic factors that influence successful cessation of illicit opioids and prescription opioid misuse may be due to the

challenge of assembling sufficiently large cohorts with the requisite information about opioid use and cessation, high relapse rate, and non-standard definitions of cessation[107,108].

Here, we present results from the first genome-wide association study (GWAS) for opioid cessation among persons meeting DSM-5 criteria for OUD[21] in a sample of African Americans (AAs) and European Americans (EAs) recruited for genetic studies of substance use disorders (SUDs)[109,110]. We also identified genetic overlap between opioid cessation and other SUDs-related traits in the UK Biobank (UKBB)[111].

Material and Methods

Participants and Diagnostic Procedures

Participants for this study were ascertained from two sources. The Yale-Penn sample includes 6,188 African American (AA) and 6,835 European ancestry (EA) participants who were enrolled in genetic studies of dependence on opioids, cocaine or alcohol between 2000 and 2017 through treatment clinics at the University of Connecticut Health Center, Yale University School of Medicine, the Medical University of South Carolina, University of Pennsylvania, and McLean Hospital in Belmont, Massachusetts[112,113]. This cohort included affected sibling pairs and additional family members, as well as unrelated cases and controls. Probands with a schizophrenia or schizoaffective disorder were excluded[112,113]. Psychiatric interviews using a computerized version of the Semi-

Structured Assessment for Drug Dependence and Alcoholism (SSADDA)[114,115] were administered to all participants and diagnoses of dependence on various substances and other psychiatric disorders were derived according to DSM criteria. A second sample of EAs was derived from the Comorbidity and Trauma Study (CATS) which has been described previously. [116] In brief, opioid dependent (OD) cases ages 18 or older were recruited from opioid substitution therapy (OST) clinics in metropolitan Sydney, Australia. Persons who had recent suicidal intent or psychosis were excluded. Controls were recruited from neighborhoods geographically proximal to the OSTs and excluded persons who used opioids recreationally more than five times (i.e., only OD cases were included in this analysis). Participants were interviewed using the SSADDA – Australia (SSADDA-OZ)[117] which was used to derive DSM-IV SUD diagnoses. Institutional review board approval for this study was obtained at all participating institutions, and written informed consent was obtained from all study participants.

Cessation Phenotype Definition

Participants who were eligible for this analysis met at least two DSM-5 OUD criteria, corresponding to a lifetime diagnosis of OUD. Current opioid cessation status was determined based on the response to the SSADDA question, “When was the last time you used an opioid drug (including illicit methadone)”. We classified individuals who last used an opioid >1 year before the date of interview as having successfully ceased opioid use and those whose

last use of an opioid was <6 months before the interview date as not ceased opioid use. Other opioid users whose self-reported last use was between 6 month and a year ago were excluded from the analysis. The ascertainment scheme and filtering steps to arrive at the sample of 1,130 AAs and 1,859 EAs informative for the opioid cessation study are shown in Figure 2.1. Persons in the CATS dataset with a lifetime diagnosis of OD were classified as ceased if the last use of an opioid was at least one year before the age at the last interview (n=337) or non-ceased if the age of last use of an opioid was the same as the age at the last interview (n=723). OD cases in the CATS sample who last used an opioid exactly one before the last interview were excluded. Characteristics of all individuals included in the GWAS are shown in Table 2.1.

Figure 2.1. Derivation of opioid cessation study samples from the Yale-Penn dataset.

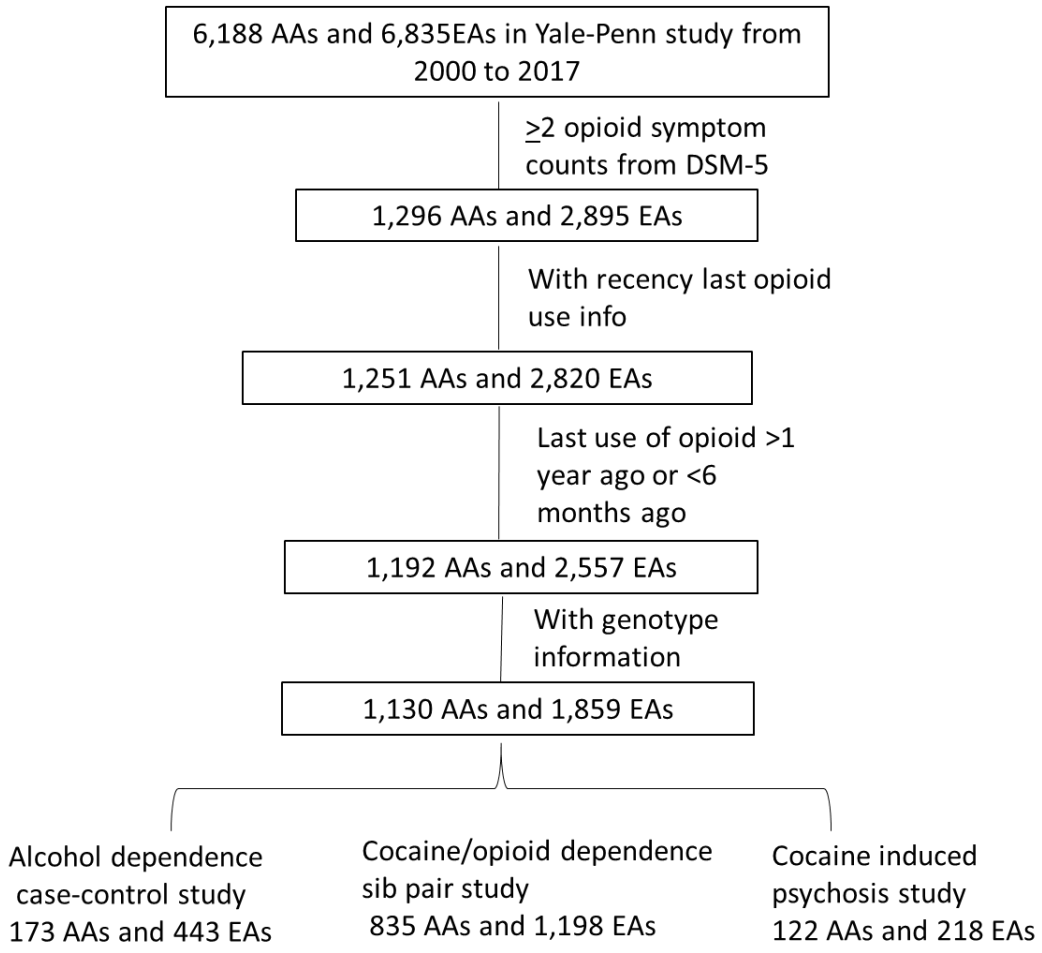


Table 2.1. Characteristics of the participants for the opioid cessation GWAS

	African Americans			European ancestry (Yale-Penn)			European ancestry (CATS)		
	Time since last use *			Time since last use *			Time since last use **		
	Not ceased	Ceased	Not classified[$]	Not ceased	Ceased	Not classified[$]	Not ceased	Ceased	Not classified[$]
Total N (% female)	682 (32.3)	448 (40.6)	101 (30.7)	1,235 (33.3)	624 (41.3)	344 (37.5)	723 (37.1)	337 (41.8)	166 (48.2)
Age (SD) in years	41.5 (8.4)	45.2 (8.4)	43.7 (9.1)	30.1 (10.2)	40.3 10.7)	35.6 (9.7)	35.9 (8.3)	39.0 (8.3)	35.5 (8.8)
OUD Symptom Counts (SD)	7.5 (2.2)	7.3 (2.6)	8.1 (2.2)	8.5 (2)	8.3 (2.1)	8.6 (2.1)	9.0 (1.4)	8.8 (1.5)	9.0 (1.4)
# of families (N in families)	42 (91)	21 (42)	2 (4)	70 (142)	20 (40)	9 (20)	N/A	N/A	N/A

[$] not classified indicated people who were from the available datasets but were not included in the current GWAS study.
* not ceased : < 6 months; ceased: > 1 year; Not classified: 6 months ≤ time since last use ≤ 1 year
** not ceased : <1 year; ceased: > 1 year; Not classified: time since last use = 1 year

Genotyping, Imputation and Quality Control

DNA specimens in the Yale-Penn sample were genotyped using the Illumina HumanOmni1-Quad v1.0 microarray (OMNI) which contains 988,306 autosomal SNPs, the Illumina Infinium Human Core Exome microarray (HCE) which contains 265,919 exome-focused SNPs and approximately 240,000 tagging SNPs to allow genome-wide imputation, or the Illumina Multi-ethnic Global Array (MEGA) which contains 1,779,819 markers that have appreciable frequency in at least one of five major populations in order to maximize genome-wide imputation accuracy. Genotyping was performed at the Yale Center for Genome analysis except for a group of participants (822 AAs and 955 EAs) who were genotyped using the OMNI1 array at the Center for Inherited Disease Research (CIDR). DNA specimens from the CATS sample were genotyped at CIDR using the Illumina Human660W-Quad BeadChip. Quality control of genotype data was performed as previously described[39]. Briefly, individuals with a call rate <98% and variants with minor allele frequency (MAF) <1% were excluded. Pairwise identity-by-decent (IBD) was calculated with PLINK[118] to determine genetic relatedness among individuals in the sample. Individuals with a pairwise IBD estimate > 25% were assigned to the same family. Self-reported males with X chromosome heterozygosity >20% and self-reported females with X chromosome heterozygosity < 20% were excluded. Population substructure in the entire sample was evaluated by analysis of principle components (PC) of ancestry using Eigensoft[119] and the multi ethnic 1000 Genome reference panel

for comparison. Individuals were classified as AA or EA according to the reference panel population to which they more closely matched. SNP genotype imputation was performed separately in AAs and EAs using the March 2012 1000 Genomes reference panel (1000 Genomes Project, 2012; http://www.1000genomes.org/) and IMPUTE2[120] implemented on the Michigan imputation server (https://imputationserver.sph.umich.edu).

Genetic Association Analysis

Association of opioid cessation with dosage of the minor allele of each SNP was evaluated using logistic regression models solved with generalized estimating equations (GEE) to correct for correlations among related individuals and included terms for age, sex and the first five PCs to correct for population sub-structure. Association testing was restricted to SNPs with an imputation score >0.8 and MAF >3%. Association tests were performed separately within each ethnic group and within each genotyping platform to account for batch effects. To prevent low frequency SNPs from inflating test statistics, a filter based on the effective number of minor alleles in cases (successful cessation) was used. SNPs with an effective number of minor alleles of fewer than 10 (N_{eff} = 2 × $N_{\text{achieving successful cessation}}$ × imputation quality × MAF) were excluded. A cutoff of 10 was chosen to control for type I error. For the CATS dataset, we obtained summary statistics of SNPs with MAF > 0.01 and imputation R^2 > 0.3, logistic association was performed using PLINK v.1.9 [121,122] adjusting for sex, age, and the first three PCs. We used a less stringent SNP inclusion criteria on CATS

dataset to maximize the variant overlap to our discovery dataset given CATS variants were genotyped on a different chip. Results were corrected for genomic inflation (λ) and combined across ethnic and batch groups via inverse variance weighted meta-analysis implemented in the program METAL[123]. The genome-wide significance (GWS) threshold was set at $p< 5.0\times10^{-8}$.

Power

Power for detecting significant association with opioid cessation in the study sample was evaluated, assuming $\alpha=5\times10^{-8}$, power $(1-\beta)=0.8$ and an additive model, separately for AAs and EAs using a genetic power calculator[124]. These analyses indicated sufficient power to detect a variant with a MAF of 0.04 and heterozygous genotype relative risks of 2.00 in AAs and 1.62 in EAs, or a variant with a MAF of 0.1 and heterozygous genotype relative risks of 1.70 in AAs and 1.40 EAs.

Assessment of SNP Effects on Gene Expression

SNPs that surpassed a threshold of $p<1\times10^{-5}$ in the opioid cessation GWAS were assessed for their potential to affect gene expression using information in the Genotype Tissue Expression Portal (GTEx)[125] (http://www.gtexportal.org) and Braineac (http://www.braineac.org/)[126] databases. GTEx contains information that links SNP genotype to expression in multiple human tissues, whereas Braineac incorporates expression data for multiple brain regions derived from 130 individuals that were obtained from the UK Brain

Expression Consortium (UKBEC) and contains information of SNPs that can affect gene expression for multiple brain regions.

Polygenic Risk Score (PRS) Analysis

We generated PRS using the summary statistics for DSM-5 symptoms counts for opioid, cocaine and alcohol dependence among individuals who were exposed to the particular substance at least once from the combined Yale-Penn and CATS datasets. These PRS were regressed on opioid cessation status of Yale-Penn participants. Genome-wide association tests for these traits were performed using the same methods as described for the opioid cessation GWAS. Three significance thresholds ($p<0.01$, 0.001 and 0.0001) were used to identify the set of SNPs explaining the greatest proportion of the total trait variance within each ancestry group according to Nagelkerke's R^2 value using the PredictABEL R package[127]. For each threshold, correlated variants were removed by LD pruning ($r^2>0.2$) to avoid over-estimating the effect of variants in large genes. Weighted PRS were constructed using the allele dosage of the variants multiplied by their effect sizes. Opioid cessation of each ancestral group was then regressed on opioid, cocaine or alcohol use disorder severity PRS calculated in that ancestral group. Results were corrected for multiple testing using the Bonferroni method.

Genetic overlap between opioid cessation and several OUD-related traits was also evaluated in the UKBB[111]. Relevant outcomes in the UKBB included "former drinker" or "ex-smoker" which reflects ability to cease use of other

substances, and “back pain for more than 3 months” which might indicate the use of illicit opioids for analgesic effect. “Back pain for more than 3 months” was chosen instead of other pain-related traits because back pain is the most common form of pain and also a leading cause of disability[128,129]. We used PRSice[130] to calculate PRS for these traits in order to estimate their genetic correlation with opioid cessation. These analyses were conducted for EAs only because the majority of participants in the UKBB are EAs. PRSice identifies the optimal p-value threshold that explains most of the trait variance for constructing PRS. PRSice models assume an additive effect for each SNP, and SNPs in high linkage disequilibrium (LD, r^2>0.25) were removed using 200 bp sliding windows[130]. PRS were calculated by summing the product of the risk allele count reported in UKBB GWAS from opioid cessation samples multiplied by the effect reported in the UKBB GWAS. Results were corrected for multiple testing using the Bonferroni method.

Pathway Analysis

Population-specific GWAS summary statistics were assessed for enrichment of functionally related genes using MAGENTA[131]. MAGENTA combines individual variant association p-values into a gene score and computes a p value that is corrected for gene size, the number of variants, and LD. A nominal p-value is then calculated for each gene set, defined as the fraction of randomly sampled gene sets of identical size less than either the 95th or 75th percentile of gene score p values, after multiple testing correction. We used the

75th percentile cutoff for significance since it demonstrates greater power for highly polygenic traits with weak effect associations[131].

RESULTS

In the Yale-Penn dataset, females had increased odds of cessation in AAs (OR=1.33, $p=1.40\times10^{-2}$) and EAs (OR=1.27, $p=1.84\times10^{-3}$). Age was associated with increased odds of longer cessation in AAs (OR=1.05 per year, $P=8.34\times10^{-12}$) and EAs (OR=1.05 per year, $p=2.00\times10^{-20}$). In the CATS dataset, persons who ceased using opioids were 4.7 years older than those who did not ($p=4.49\times10^{-4}$). Although individuals with mild OUD were included, the mean number of OUD symptoms counts among participants in both datasets corresponds to a severe diagnosis.

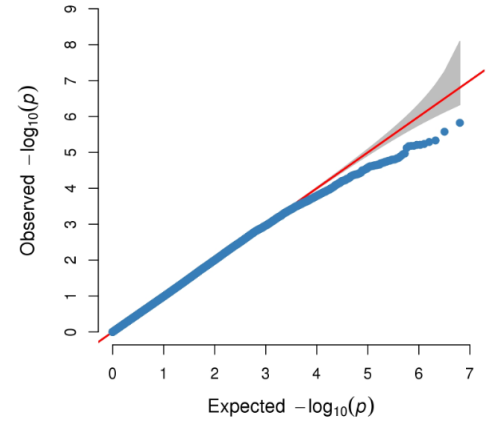
Genetic association findings for opioid cessation

There was no evidence of p-value inflation in either population group (Figure 2.2). Although no SNPs reached genome-wide significance (Figure 2.3), SNPs showed suggestive evidence of association ($p<1\times10^{-5}$) with opioid cessation were located in nine independent regions including three specific to EAs and six in the combined sample (Table 2.2). Several of these associations were supported by evidence with surrounding SNPs (Figure 2.4).

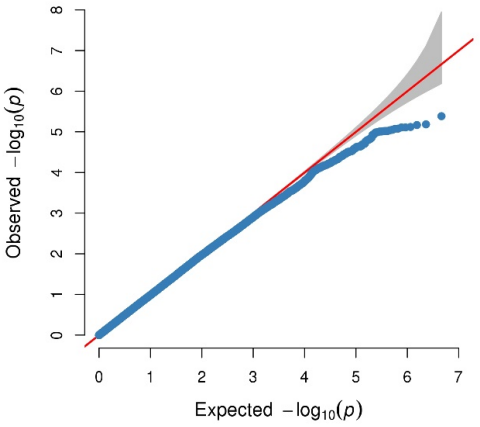
Figure 2.2. QQ plot of opioid cessation meta-analysis results.

(A) AAs (Yale-Penn only), (B) EAs only (Yale-Penn +CATS), and (C) AAs and EAs combined. The genomic inflation factors for African Americans is 0.985, for European Americans is 0.981, and for combined is 0.991.

(A)



(B)



(C)

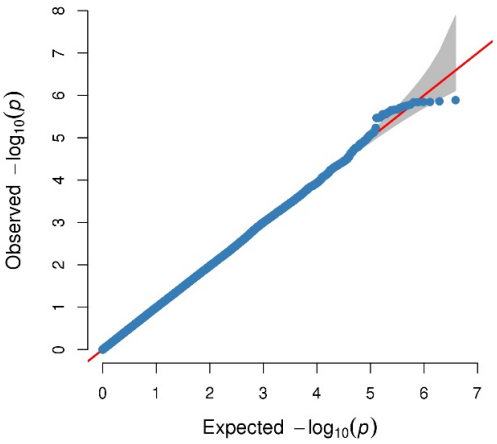


Figure 2.3. Manhattan plot of opioid cessation meta-analysis results.

(A) AAs (Yale-Penn only), (B) EAs only (Yale-Penn +CATS), and (C) AAs and EAs combined. Red line indicates genome-wide significant threshold ($p = 5 \times 10^{-8}$), green line indicates sub genome-wide significant threshold ($p = 1 \times 10^{-6}$).

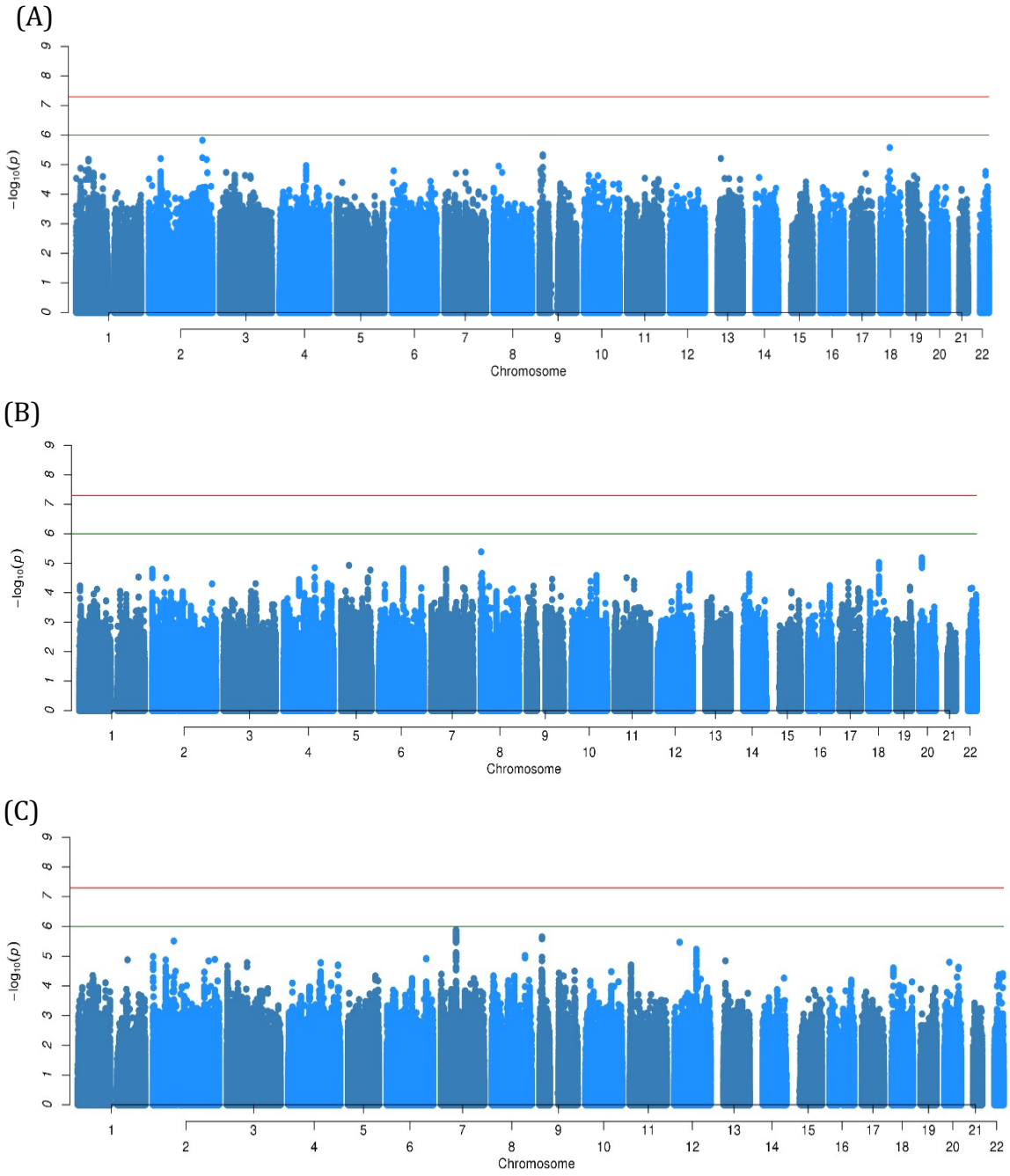


Table 2.2. Association findings for opioid cessation ($p< 1\times 10^{-5}$).

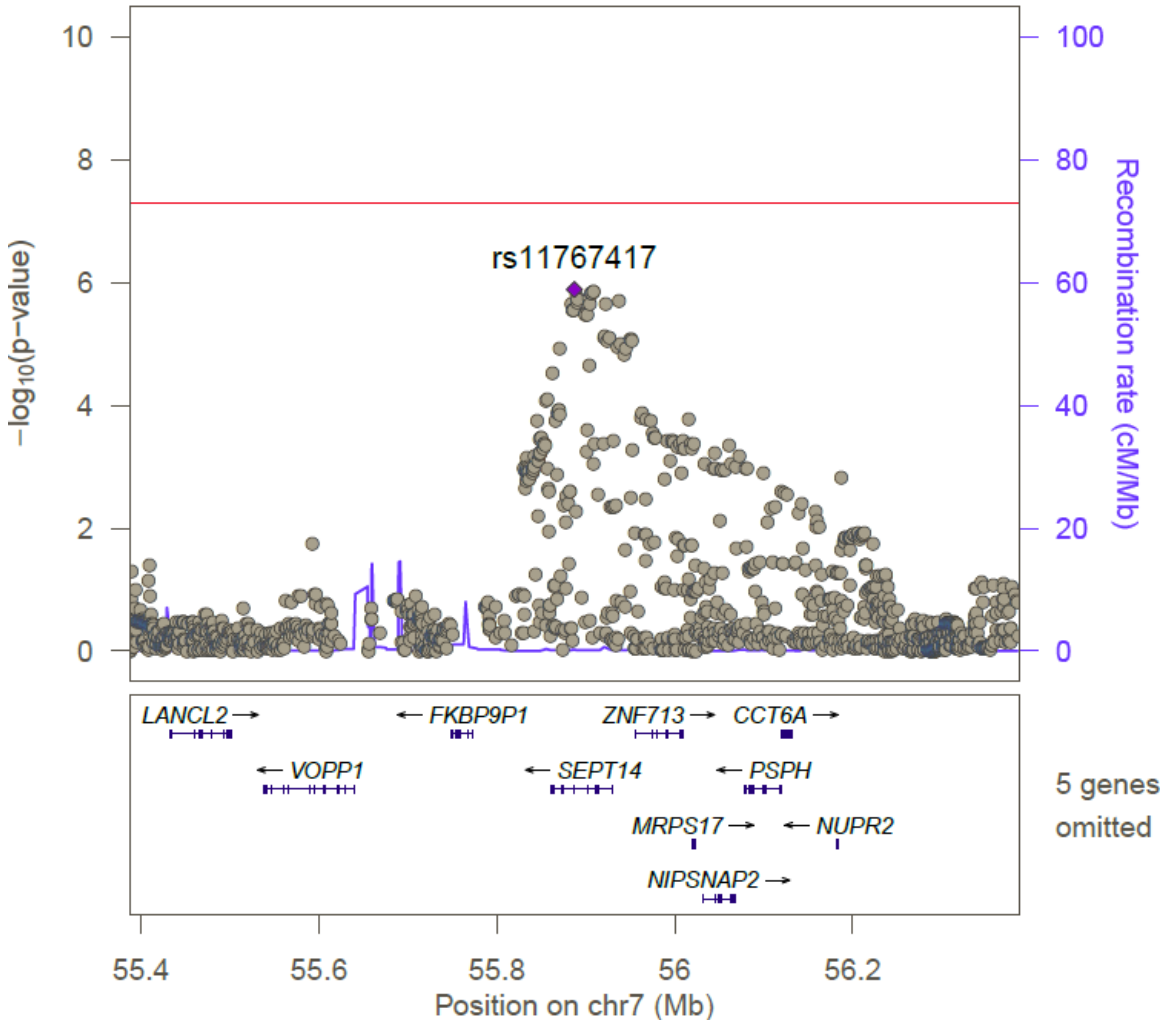
Location	Gene	African Americans (AAs)			European Americans (EAs)			AAs+EAs	
		OR	P	Dir	OR	P	Dir	OR	P
7:55887138	SEPT14	1.29	1.06E-02	+++	1.50	2.02E-05	++-+	1.39	1.30E-06
9:9783693	PTPRD	1.48	4.33E-04	++?	1.25	6.62E-04	++++	1.31	2.24E-06
2:76417353	SUCLA2P2 / AC073091.2	0.69	2.20E-04	---	0.73	4.11E-03	----	0.71	3.09E-06
12:19767596	AEBP2	1.36	7.53E-03	+++	1.50	1.41E-04	+?++	1.43	3.37E-06
8:2082245	MYOM2	0.93	4.53E-01	+--	1.72	4.11E-06	++++	1.22	7.77E-02
12:81223948	MIR617	1.43	5.37E-03	++?	1.48	2.88E-04	++++	1.46	5.84E-06
20:10086110	SNAP25-AS1	1.11	5.48E-01	+??	1.44	6.53E-06	+-++	1.24	3.23E-03
8:121847800	RP11-713M15.2 / RP11-369K17.1	1.27	4.76E-02	++-	1.27	7.45E-05	++++	1.27	9.47E-06
18:40805792	RIT2-SYT4	1.28	3.03E-01	+??	1.74	9.50E-06	++++	1.61	3.22E-04

Dir: effect direction in the OMNI, HCE and MEGA array samples in Yale-Penn dataset and in the CATS dataset, respectively.

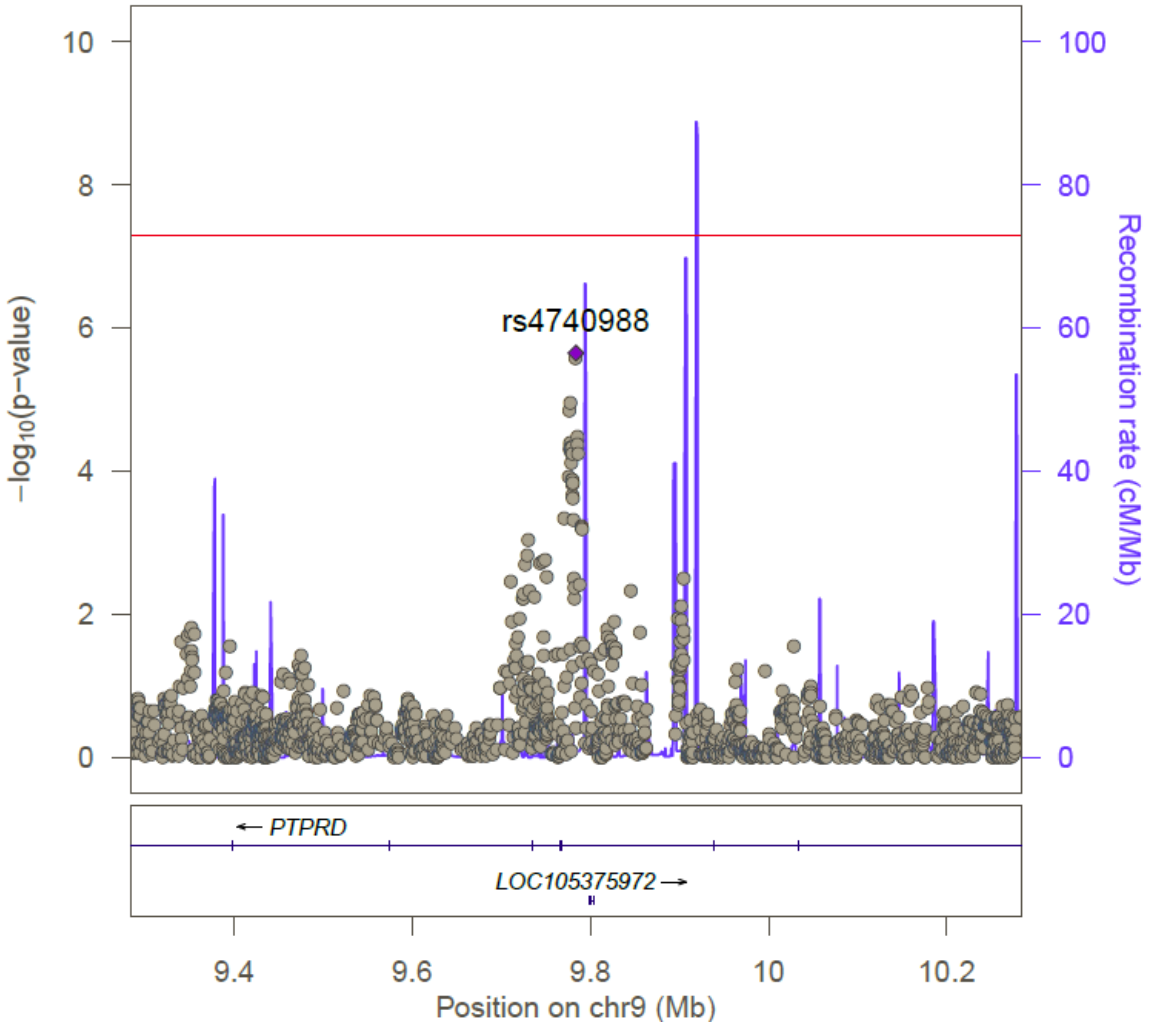
Figure.2.4. Locuszoom plot of association results.

Linkage disequilibrium and recombination rate around the four biologically relevant loci (p <1×10-5) from meta-analysis of EA+AA datasets (panels A and B) or EA datasets only (panels C and D). The red line marks the genome-wide significant threshold. The LD estimate between top variants and its surrounding variants are indicated by color on the top left corner, where recombination rates were indicated by blue horizontal lines. The LD estimates in A and B are not present given the admixture population. Arrows on the horizontal blue lines show the direction of transcription, rectangles are exons.

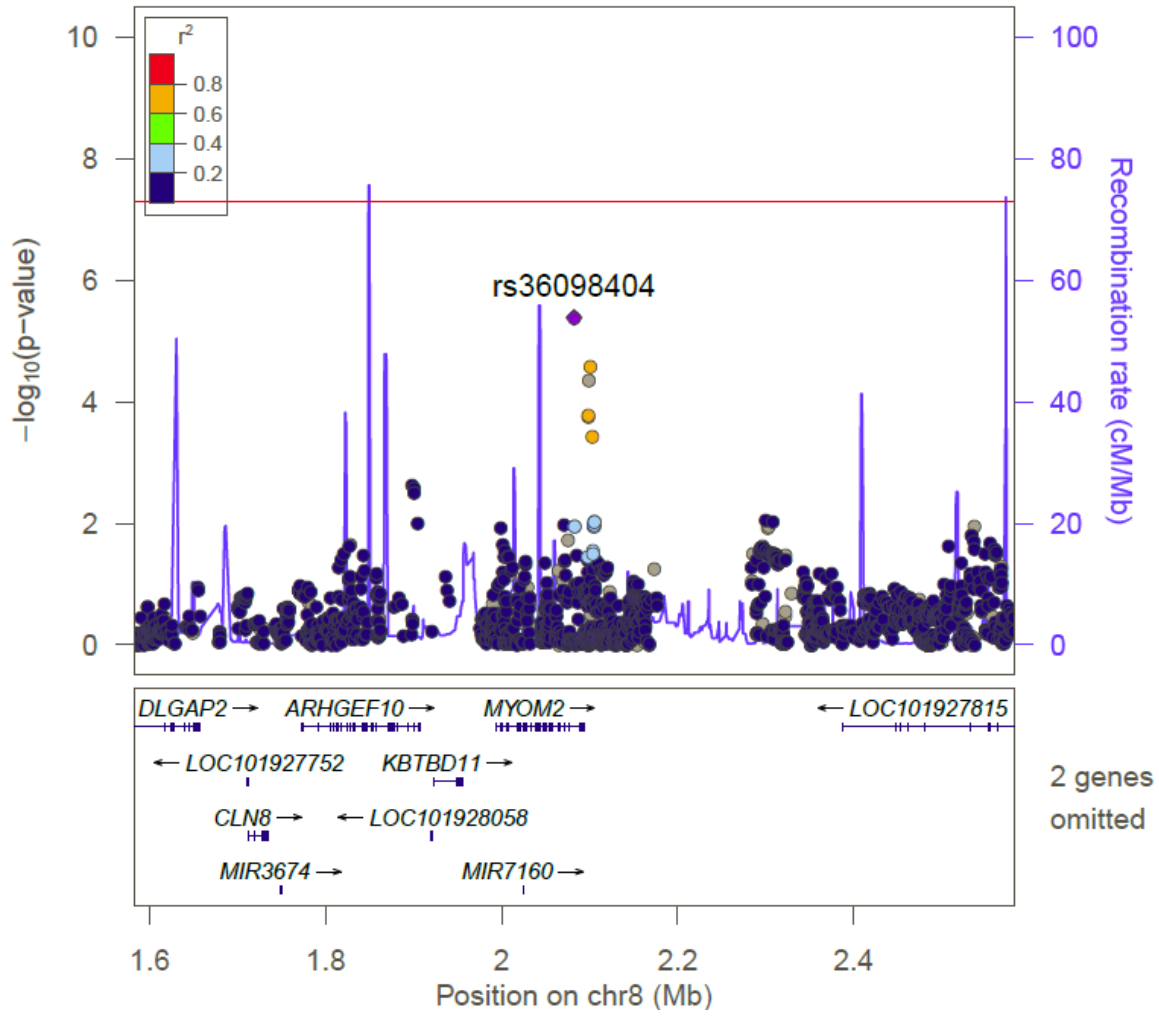
(A)


rs11767417
−log10(p−value)
Recombination rate (cM/Mb)
10
8
6
4
2
0
100
80
60
40
20
0
LANCL2
FKBP9P1
ZNF713
CCT6A
VOPP1
SEPT14
PSPH
MRPS17
NUPR2
NIPSNAP2
5 genes
omitted
55.4
55.6
55.8
56
56.2
Position on chr7 (Mb)

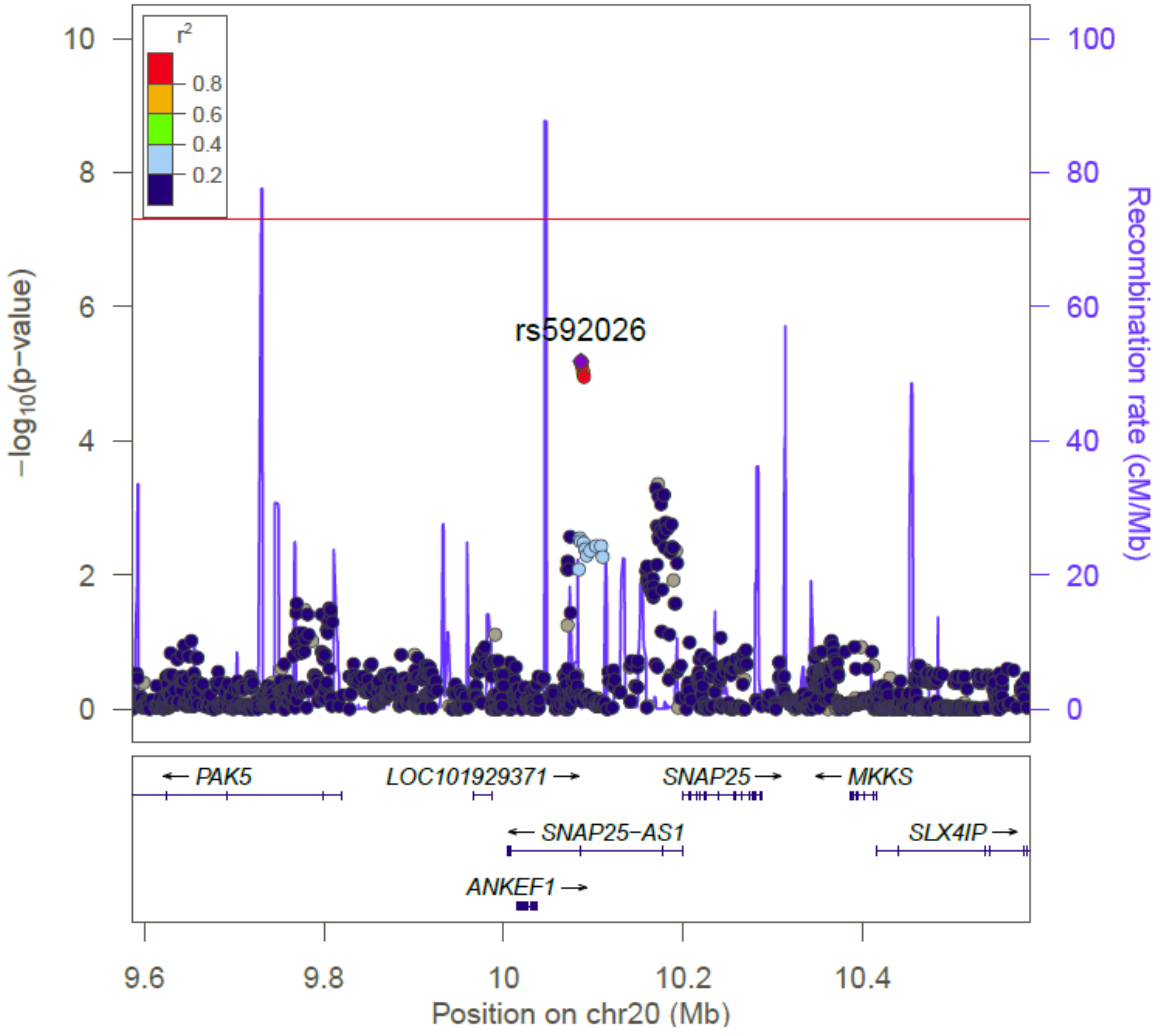
(B)


10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
rs4740988
PTPRD
LOC105375972
9.4
9.6
9.8
10
10.2
Position on chr9 (Mb)

(C)


r²
0.8
0.6
0.4
0.2
rs36098404
−log10(p−value)
Recombination rate (cM/Mb)
DLGAP2
ARHGEF10
MYOM2
LOC101927815
LOC101927752
KBTBD11
CLN8
LOC101928058
MIR3674
MIR7160
2 genes omitted
Position on chr8 (Mb)

(D)


r²
0.8
0.6
0.4
0.2
rs592026
−log10(p−value)
Recombination rate (cM/Mb)
PAK5
LOC101929371
SNAP25
MKKS
SNAP25−AS1
SLX4IP
ANKEF1
Position on chr20 (Mb)

Biological pathways and gene sets that influence opioid cessation

After multiple test correction, significantly enriched pathways were observed in EAs for vitamin D metabolism (p_{adj}= 3.79 ×10^{-2}) containing nine genes while only three were expected (Table 2.3A) and fibroblast growth factor (FGF) signaling (p_{adj}= 2.39 ×10^{-2}) containing 20 genes while only 10 were expected (Table 2.3B). No pathways were significantly enriched in AAs.

Table 2.3 Genes in significantly enriched pathways from MAGENTA in European Americans.

(A)Vitamin D metabolism and pathway

Gene	Chr	Start Pos.	End Pos.	Size (kb)	Number of SNPs	Number of independent SNPs	Gene p-value	Top-ranked SNP			
								ID	Z-score	p-value	Odds Ratio
COX15	10	101468504	101492423	24	197	5	1.14E-01	rs11190244	2.59	9.63E-03	0.85
CYP2D6	22	42522500	42526883	4	217	8	2.18E-01	rs742086	2.33	1.98E-02	1.18
CYP24A1	20	52769987	52790516	21	261	28	2.02E-01	rs1298513	2.88	3.92E-03	0.84
CYP27B1	12	58156116	58160976	5	79	7	1.07E-01	rs182005106	2.37	1.79E-02	1.29
POR	7	75544419	75616173	72	238	13	2.35E-02	rs11311523	3.16	1.56E-03	1.23
RARA	17	38465422	38513895	48	15	11	1.30E-01	rs59214602	2.25	2.42E-02	1.20
RARB	3	25469753	25639422	170	607	56	2.49E-01	rs17016192	2.87	3.77E-03	0.72
RARG	12	53604352	53626036	22	65	11	1.23E-01	rs73309166	2.44	1.48E-02	1.36

(B) FGF signaling pathway.

Gene	Chr	Start Pos.	End Pos.	Size (kb)	Number of SNPs	Number of independent SNPs	Gene p-value	Top-ranked SNP			
								ID	Z-score	p-value	Odds Ratio
FGF2	4	123747862	123819390	72	255	15	1.05E-03	rs35287534	3.77	1.62E-04	0.76
FGF3	11	69624735	69634192	9	203	33	1.15E-01	rs10908251	2.89	3.89E-03	0.78
FGF4	11	69587796	69590171	2	249	36	1.66E-01	rs10908251	2.89	3.89E-03	0.78
FGF5	4	81187741	81212171	24	279	21	1.72E-01	rs3796594	2.74	6.06E-03	0.74
FGF6	12	4543307	4554780	11	177	19	1.37E-01	rs2244388	2.76	5.84E-03	0.83
FGF7	15	49715374	49779523	64	361	10	1.74E-01	rs1583060	2.62	8.93E-03	1.34
FGF12	3	191857181	192445388	588	887	71	1.76E-01	rs2221465	3.17	1.52E-03	1.24
FGFR1	8	38268655	38326352	58	205	16	1.10E-01	rs10637920	2.67	7.03E-03	1.24
FGFR3	4	1795038	1810599	16	103	7	6.72E-02	rs712983	2.65	8.13E-03	0.82
FGFR2	10	123237843	123357972	120	366	42	5.64E-02	rs10886938	3.20	1.38E-03	0.79
PPP2R1B	11	111597631	111637169	40	89	5	1.41E-01	rs61899413	2.27	2.34E-02	1.21
PPP2R2C	4	6322304	6474326	152	416	56	2.30E-01	rs113823191	2.87	4.06E-03	0.72
PPP2R5A	1	212458878	212535205	76	371	21	2.30E-01	rs11405845	2.57	1.02E-02	1.20

PPP2R5B	11	64692142	64701950	10	40	12	2.02E-01	rs56339918	2.24	2.55E-02	1.15
PPP2R5D	6	42952329	42980080	28	257	7	4.83E-02	rs3805946	2.94	3.30E-03	1.19
MAP3K6	1	27681669	27693337	12	79	10	8.97E-02	rs4246507	2.56	1.05E-02	1.19
SPRY1	4	124317955	124324909	7	189	24	1.38E-02	rs300576	3.35	8.03E-04	1.37
SPRY2	13	80910111	80915086	5	129	11	8.17E-02	rs9601380	2.68	7.29E-03	0.82
MRPL38	17	73894723	73901181	6	199	11	1.18E-01	rs142057056	2.67	7.66E-03	1.50
PEBP4	8	22570764	22785421	215	366	57	9.68E-02	rs200317753	3.05	2.27E-03	0.76

Genetic correlation with other SUD traits

Table 2.4 shows that the PRS for severity of cocaine use disorder was a significant predictor of opioid cessation in AAs ($p= 9.0\times10^{-4}$), whereas the PRS for severity of alcohol use disorder was a significant predictor of opioid cessation in EAs ($p =3.00\times10^{-4}$). The PRS for severity of opioid use disorder was not associated with opioid cessation in either AAs or EAs. Among the traits considered in the UKBB dataset, the PRS for back pain persistent for more than 3 months was significantly associated with cessation in EAs ($p=1.25\times10^{-7}$). Opioid cessation was also associated with PRSs for being a former drinker ($p=6.24\times10^{-4}$) and ex-smoker ($p=4.44\times10^{-3}$). However, the proportion of variance for opioid cessation explained by the trait PRS were small, ranging from R^2= 0.00747 for being an ex-smoker and R^2= 0.0211 for having back pain for more than 3 months in EAs.

Table 2.4. Association and model fit between polygenic risk scores of traits related to opioid cessation in (A) African Americans and (B) European Americans.

(A)

	Trait	GWAS N	GWAS p threshold	PRS P_{adj}	R^2
Yale-Penn dataset	Alcohol severity score	5,561	0.01	NS	4.50E-03
	Cocaine severity score	4,015	0.001	9.00E-04	1.86E-02
	Opioid severity score	2,275	0.01	NS	3.30E-03

(B)

	Trait	GWAS N	GWAS p threshold	PRS P_{adj}	R^2
Yale-Penn + CATS dataset	Alcohol severity score	7,655	0.01	3.00E-04	1.49E-02
	Cocaine severity score	5,248	0.01	NS	4.50E-03
	Opioid severity score	4,701	0.001	NS	6.00E-04
UKBB dataset	Ex-smoker	83,133	1	4.44E-03	7.47E-03
	Former drinker	21,894	0.09185	6.24E-04	9.93E-03
	Back pain for >3 months	84,489	0.1822	1.25E-07	2.11E-02

GWAS N: the sample size of the trait where the variants from the summary statistics were obtained.
GWAS p threshold: the p value threshold for including a variant from the trait summary statistics for PRS construction.
PRS P_{adj}: the adjusted p-value for the trait PRS in a prediction model for opioid cessation. NS: not significant.
R^2 : the variance of opioid cessation explained by the trait PRS.

Discussion

To our knowledge, this is the first GWAS for opioid cessation. Although no genome-wide significant associations were identified in the EA group or in the combined sample of AAs and EAs, we obtained suggestive evidence for association of cessation from opioid use for at least one year among persons who were diagnosed with OUD with variants in several genes not previously linked to OUD. Pathway analysis that was seeded with summary GWAS information implicated two biological processes in successful opioid abstinence including vitamin D metabolism and FGF signaling. Analyses of PRSs computed for risk of other SUDs, ability to quit smoking or drinking, and persistent back pain suggest shared a genetic underpinning of these traits with opioid cessation.

Several of the top-ranked genes associated with opioid cessation have been implicated previously in substance use disorders or other psychiatric traits. Protein tyrosine phosphatase receptor type D *(PTPRD)* is abundantly expressed in CNS ventral midbrain neurons involved with reward, locomotor, and sleep processes in both mice and humans[110,132], it is an important regulator of axon growth[133]. Myomesin 2 (*MYOM2)* encodes for a major protein in the muscular tissue, *MYOM2* is nominally associated with nicotine and alcohol dependence in Australian and Dutch populations[134]. Synaptosomal-associated protein of 25kDa *(SNAP-25)* controls the release of neurotransmitters by modulating voltage-gated calcium channels[135]. Studies have linked *SNAP-25* to response to antipsychotic treatment[136], morphine-associated contextual memory retrieval[137], and the risk of

ADHD[138]. The top variant rs592026 (Supplemental Figure 4D) in *SNAP-25-AS1* is located 113kb upstream of *SNAP-25* and is also an eQTL for in *SNAP-25* in thalamus ($P = 1.8\times10^{-4}$).

A role for vitamin D metabolism and FGF signaling pathways identified from biological pathway analyses that relied on evidence from GWAS in addiction relevant biology is supported by previous research. A neuroprotective effect of vitamin D against methamphetamine-induced dopamine depletion has been demonstrated.[139] Vitamin D3 has been shown to reduce neuropathic pain by modulating opioid signaling in the rodent brain[140]. Low vitamin D level was observed in methadone maintenance patients[141] and associated with a higher dose opioid for treatment of cancer patients.[142] In addition, vitamin D inadequacy results in worse physical functioning and poorer health perception.[143] One explanation for this observation is that individuals who cease opioid use might be more physically fit because of better vitamin D metabolism compared to persistent opioid users. Several other genes in the vitamin D metabolism pathway have been linked to opioids. *CYP2D6* ($P_{best\_SNP} = 0.02$) is a member of the P450 family and encodes an enzyme that has many opioids substrates[144]. Polymorphisms in *CYP2D6* can dramatically affect the metabolic capacity leading to under- or over-exposure to opioids[145]. *CYP2D6* ultra-rapid metabolizers have a higher chance to experience adverse events with opioid prescriptions[146]. In the presence of opioids, *RARA* ($P_{best\_SNP} = 0.02$) forms a complex with *RXRA*, which activates mu-opioid receptor and modulates drug–seeking behaviors[147]. Some

mutations in the vitamin D receptor gene (*VDR,* P_{gene} = 0.04, P_{best_SNP} = 0.001) have been associated with impulsivity in the context of alcohol dependence[148].

The FGF pathway has been previously implicated in opioid metabolism. One study showed that FGF is involved in the development of analgesic tolerance to the opioid agonist morphine.[149] In addition, the FGF receptor was identified as a converging site between mu-opioid receptor and growth factor signaling pathways[150]. One of the significantly enriched genes identified from this pathway, *MAP3K6* (P_{gene} = 0.09, P_{best_SNP} = 0.01), is differentially expressed in striatum and prefrontal cortex among chronic and acute morphine treated mice[151]. There are also links between vitamin D metabolism and FGF signaling. *FGF23*, a fibroblast growth factors that is highly expressed in bones, reduces levels of activated vitamin D[152] and thus leads to reduced bone mineral density and a higher risk of fracture[153]. Interestingly, long term use of opioid analgesics has been associated with bone loss[154].

PRS analysis revealed genetic overlap between opioid cessation and several other traits. Although the observed associations of opioid cessation with PRS for cocaine and alcohol symptom counts may be due to the ascertainment of our sample which is enriched for individuals dependent on cocaine and alcohol, shared genetic underpinnings between opioid cessation and risk of other SUDs is supported by studies in twins[155] and unrelated individuals[156]. It is not surprising that chronic pain is genetically correlated with increased substance use among patients on methadone treatment [157]. However, randomized clinical

trials suggest that patients with chronic pain do not receive a benefit from opioid treatment, but suffer from more opioid therapy adverse events[128,158]. The limited pain reducing property of opioids might explain the positive association between genetic predispositions to having chronic pain and opioid cessation. In addition, the Netrin-1 receptor gene *DCC* was found to be related to chronic opioid exposure and chronic back pain in both mice and human studies[159,160].

Several limitations of our study should be noted. First, we used cross sectional data to study a phenotype that would require long-term follow up to more accurately define cessation. It is possible that some people who ceased opioids at the time of interview became users again. Second, we used a slightly different definition for cessation in the CATS dataset because these participants were recruited with multiple versions of the psychiatric interview instrument that resulted in a time frame of last opioid use that was less precise than that for subjects in the Yale-Penn dataset. This might limit the utility of CATS for replication. Third, several PRSs were constructed on phenotypes from Yale-Penn and CATS dataset, which has a decent amount of sample overlap to sample used for opioid cessation, this might affect the validity of the PRS results. Fourth, the opioid cessation GWAS sample had limited power for detecting robust association signals. For example, a sample of 300 additional cases would be needed to attain 80% of power for detecting genome wide significance with the top-ranked variant rs11767417 if the observed effect size is the true effect size. However, although only sub-genome wide significant associations were

identified, we observed supportive evidence for association of the top variants with adjacent SNPs (Supplemental Figure 4). Future studies with adequately powered samples should be conducted to validate some of our findings. Finally, although small, the proportion of variance for opioid cessation explained by the significant trait PRS is in the range of similar studies[161].

Conclusion

Although no genome-wide significant variants were found in this GWAS for opioid cessation, the top-ranked variants are located in genes previously linked to SUD traits that have important roles in neural growth and signal transduction in brain regions related to pain and reward systems. Pathways derived from genetic evidence suggest additional biological mechanisms that influence opioid cessation.

CHAPTER THREE: Identifying factors associated with opioid cessation in African Americans and European Americans using machine learning

Abstract

Background and Aims Opioid use disorder patients (OUD) who stop using opioids may or may not use pharmacological treatment. Racial disparities in OUD management exist, however, and there is limited research on factors that influence opioid cessation in different population groups.

Methods: We employed multiple machine learning prediction algorithms (LASSO, random forest, deep neural network, and support vector machine) to assess factors associated with ceasing opioid use in a sample of 1,192 African Americans (AAs) and 2,557 European Americans (EAs) who met DSM-5 criteria for OUD. Values for nearly 4,000 variables reflecting demographics, alcohol and other drug use, general health, non-drug use behaviors, and diagnoses for other psychiatric disorders, were obtained for each participant from the SSADDA, a detailed semi-structured interview.

Results: Support vector machine models performed marginally better on average than other machine learning methods with maximum prediction accuracies of 75.4% in AAs and 79.4% in EAs. Subsequent stepwise regression considered the 83 most highly ranked variables across all methods and models and identified less recent cocaine use ($p<5\times10^{-8}$), shorter duration of opioid use ($p<5\times10^{-6}$), and older age ($p<5\times10^{-9}$) as the strongest independent predictors of opioid cessation in both AAs and EAs. Attending self-help groups for OUD was

also an independent predictor ($p<0.05$) in both population groups, while less gambling severity ($P = 3.32\times10^{-2}$) was specific to AAs and PTSD recovery ($P = 7.88\times10^{-5}$), recent antisocial behaviors ($P = 2.69\times10^{-3}$), and atheism ($P = 1.34\times10^{-2}$) were specific to EAs. Factors related to drug use comprised about half of the significant independent predictors in both AAs and EAs, with other predictors related to non-drug use behaviors, psychiatric disorders, overall health, and demographics.

Conclusions: These proof-of-concept findings provide avenues for hypothesis-driven analysis, and will lead to further research on strategies to improve OUD management in EAs and AAs.

Key words: opioid use disorder, opioid cessation, opioid abstinence, machine learning, AI, diagnostic questionnaire, feature selection, outcome prediction

Introduction

Use of illicit and misuse of prescription opioids is a significant global problem that affects the health and economic welfare of individuals, families, and society. The U.S. opioid overdose rate has quadrupled since 1991[162]. In 2017, more than 47,000 Americans died of an opioid overdose, and 36% of these deaths involved prescription opioids [100]. A major goal in treating opioid use disorder (OUD) is abstinence, or complete cessation, of opioid use, other than the use of prescribed opioid replacement therapy. There is not a single, clinically accepted definition of cessation that specifies the length of abstinence required before an individual is no longer considered to have OUD [163,164]. DSM-5 identifies sustained remission from OUD as a one-year period during which no criteria for the disorder (other than craving) are met [165].

Population differences affect multiple aspects of the current epidemic. Although opioid use nationally is higher among European-Americans (EAs) than African-Americans (AAs), the opioid death rate has increased more sharply among AAs than EAs [166]. AAs have less access to treatment for OUD [167], are less likely to obtain opioid prescriptions for pain management [168], and are incarcerated at a higher rate for illicit opioid use [169] than EAs. Previous research on OUD-related outcomes has been conducted primarily in combined ethnic groups or in EAs only [170], limiting the identification of key population differences in opioid use and treatment outcomes.

Although moderately correlated with opioid cessation, factors contribute to opioid treatment completion such as age, employment status, and age at first drug use have been identified from a mixed ethnicity sample[163]. Other factors are likely to influence cessation, such as pain experiences, general health, and the use of antidepressants [171-173]. Delineation of these factors could inform OUD treatment strategies that may differ across population groups; or could be useful for individuals with OUD who aim to reduce or stop their opioid use. However, studies thus far have tended to focus on a small number of clinically relevant factors such as the dosage, duration, and formulation of medication-assisted treatment of substance use disorders [174-176]. Large epidemiological studies of OUD [177-179] comprised of thousands of potential predictors would allow a systemic, hypothesis-free query to identify factors predicting opioid cessation.

Statistical methods are generally limited in their ability to sort through large numbers of predictors[180]. Data mining using machine learning, which is particularly well suited for identifying predictive factors among thousands of variables [181,182], has successfully identified predictor variables for a diverse set of outcomes [183-187]. Here, we applied multiple machine learning techniques to evaluate a large set of clinical, demographic, general health, and behavioral variables in a large, racially mixed cohort of individuals who were ascertained for genetic studies of substance use disorders, but not necessarily treated for OUD, to identify factors that influence opioid cessation (defined as self-reported last illicit opioid use and/or prescription opioids misuse > 1 year before the interview

date). Our study identified additional factors associated with cessation, including several that are population-specific. These findings support an individualized approach to improve the outcome of cessation attempts.

Material and Methods

Participants and Assessments

Participants for this study were selected from a cohort of 6,188 African Americans (AAs) and 6,835 European Americans (EAs) who were recruited for genetic studies of opioid, cocaine, or alcohol dependence between 2000 and 2017 through advertisements and treatment clinics at Yale University School of Medicine, the University of Connecticut Health Center, the University of Pennsylvania, the Medical University of South Carolina, and McLean Hospital [112,113]. This cohort included affected sibling pairs and additional family members, as well as unrelated cases and controls. Probands with schizophrenia or bipolar affective disorder were excluded [112,113]. Information about the use of various substances, demographics, general health, behavior, and other psychiatric illnesses was obtained by interview using the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA) [177,188]. Substance use disorder (SUD) and psychiatric disorder diagnoses were established according to DSM-IV criteria. Institutional review boards from each recruitment site and Boston University approved this study, and written informed consent was obtained from all participants.

Opioid Cessation Definition

Participants who were eligible for this analysis met at least two DSM-5 criteria for OUD, corresponding to a lifetime diagnosis of mild to severe OUD. Current opioid cessation was determined by the response to the question, “When was the last time you used an opioid drug (including illicit methadone).” This question was asked as part of a series of items asked about illicit or non-prescribed use of opioids. Individuals who last used an opioid >1 year before the date of interview were considered to have achieved cessation and those whose last use of an opioid was <6 months before the interview date were classified as non-cessation. Persons who used opioids between 6 months and 1 year before the interview date were excluded from further analysis. Filtering steps yielding a sample of 1,192 AAs and 2,557 EAs for analysis is shown in Appendix-I. Figure S1.

Phenotype Data Processing

Preprocessing of 3,956 SSADDA variables was performed prior to machine learning analyses. Variables with narrative or invariable responses, containing redundant information (e.g., specific date of different episodes, drug names), and with a response rate <90% were removed. Missing values for binary and categorical variables were recoded as indicator variables to accommodate missing responses. Missing values for continuous variables were imputed to the population group mean value. Missing values for ordinal variables related to recency of drug use were assigned the highest level indicating less recent use,

since the missing values tend to indicate “stopped using for a long time” according to other variables in the dataset. Z-score normalization (mean of 0 and variance of 1) was applied to continuous variables within each population to minimize scaling issues. The number of variables remaining after these steps was 3,315 in AAs and 3,738 in EAs.

Machine Learning Analyses.

AAs and EAs were analyzed separately based on population differences in the epidemiology of opioid use and OUD. Variables were grouped into three nested sets defining three analytical models to explore the prediction accuracy blind to the individual’s opioid or other drug use activities. This approach was adopted to enhance identification of non-drug use variables whose effects may be masked or confounded by variables related to drug use and are highly correlated with the cessation outcome. Model 1 contained all variables except those related to recency of opioid use that are strongly correlated with cessation (variable n=3,093 in AAs and n=3,503 in EAs). Model 2 further excluded all opioid-related variables (n=2,863 in AAs and n=3,252 in EAs). Model 3 further excluded all drug use variables, leaving only demographic, non-SUD diagnoses and behaviors, and other health-related variables (n=1,656 in AAs and n=1,907 in EAs). Models were evaluated using four machine learning methods described in Supplemental Materials to identify variables predicting opioid cessation. We modeled different types of intra-variable relationships between predictors and the outcome using linear (LASSO) [189] and linear support vector machine with

recursive feature elimination (SVM) [190] and non-linear (random forest with recursive feature elimination (RF) [191] and deep neural network with feature selection (DNN) [192]) techniques. These four methods were applied to capture predictive variables under different model assumptions and allow for different outcome-predictor relationships. Variables from each model that were associated with the highest accuracy reflected by either F1 score or area under the curve (AUC) and generated by each machine learning method were retained. The F1 score is a harmonic measure of precision (true positive / [true positive + false negative]) and recall (true positive / [true positive + true negative]), defined by 2*(precision * recall)/(precision + recall) at a given case/control split, and AUC is an overall evaluation of model performance that accounts for the true positive and false positive rates for all possible diagnostic splits [193,194]. Both measurements were considered because of their popularity in clinical settings [195]. The F1 score was used to assess accuracy due to limitations of the AUC, which includes bias when performed on imbalanced datasets as well as impractical and uninterpretable split points for evaluation [193,196].

Statistical Methods for Testing the Association of Opioid Cessation with Phenotypic Variables

To determine which variables selected by the machine learning methods independently predict cessation, we applied different cutoffs for the importance measurement of each machine learning method, namely the odds ratio (OR) for LASSO, coefficient[197] denoted by weight for SVM, feature importance[197] for RF,

and activation potential[192] for DNN. For LASSO, we chose variables that yielded ORs >1.05 or <0.95. We applied the following criteria for selecting variables from SVM and RF analyses depending on the number of variables (n) selected for each model: (1) if n >200, the top 30% of variables measured by absolute weight in SVM or importance in RF were designated as high impact; (2) if 100 < n < 200, the top 50% were selected; and (3) if n<100, all variables were designated as high impact. For DNN, all selected variables were designated "high impact". Joint association tests were performed using bi-directional stepwise logistic regression that included 83 "high-impact" variables that were culled from three models across four machine learning methods in the AA and EA datasets. Variables that yielded the highest Akaike information criterion (AIC) with p <0.05 from bi-directional stepwise logistic regression were grouped into "drug related", "behavioral", "general health", and "demographic" categories.

Results

Characteristics of the study samples are shown in Table 3.1. The sample included 1,069 unrelated AAs and 2,252 unrelated EAs, as well as 123 AAs and 305 EAs participants who were members of families containing a pair of siblings both with either opioid or cocaine dependence. There is a higher proportion of females among individuals who ceased opioids in both AAs (OR=1.35, P=6.7 $\times 10^{-3}$) and EAs (OR=1.31, P=1.1 $\times 10^{-3}$) compared to those who did not cease. Furthermore, participants who ceased opioid use were older by an average of 3.2 years in the AA group (P=1.0 $\times 10^{-10}$) and 6.1 years in the EA group (P=2.2 $\times 10^{-}$

[16]) than those who did not cease. The mean number of lifetime DSM-5 OUD criteria met among those who ceased opioid use was not significantly different from those who did not cease.

Table 3.1. Participant characteristics.

		Time since last use	
		≤ 6 month (not cease)	>1 year (ceased)
AAs (N=1192)	Total N (% female)	701 (33.5%)	491 (40.5%)
	Age (Mean ± SD)	42.6 ± 8.5	45.6 ± 8.3
	OUD Symptom Counts (Mean ± SD)	7.8 ± 2.4	7.6 ± 2.5
	# of families (N in families)	35 (76)	23 (47)
EAs (N=2557)	Total N (% female)	1714 (34.4%)	843 (40.6%)
	Age (Mean ± SD)	34.4 ± 10	40.5 ± 10.3
	OUD symptom Counts (Mean ± SD)	8.8 ± 1.9	8.4 ± 2.3
	# of families (N in families)	114 (241)	31 (64)

AAs: African Americans, EAs: European Americans, OUD: opioid use disorder.

Feature selection

The F1 score was generally higher across models in both AAs and EAs using SVM than the other machine learning algorithms (Appendix-I. Figure S2), although the differences in F1 score across methods were generally small, especially for models 1 and 2. A detailed discussion of the performance of each method for the three models is provided in Appendix - I.

Appendix-I. Figure S3 shows the overlap of high impact variables chosen by the four machine learning methods. LASSO “high impact” variables almost entirely overlap with those from the other methods, while DNN-selected variables overlap the least with other method-selected variables. The majority of variables selected by non-LASSO methods are unique to those methods, however, there was high overlap in “high impact” variables selected by SVM and RF. Age was among the five top-ranked variables consistently identified by each method for each model in both AAs and EAs (Appendix –I. Table S1).

Factors Associated with Opioid Cessation

Stepwise regression analysis that considered independent predictors among 83 “high impact” variables culled from all models and machine learning methods (Table 3.2) revealed that age was one of the most significant predictors of opioid cessation in both AAs (OR=2.44 per standard deviation (SD) increase in age $P=1.41\times10^{-12}$) and EAs (OR=2.00 per SD increase in age, $P=5.74\times10^{-9}$). Variables related to drug use comprised over 50% of the nominally significant predictors of opioid cessation in AAs (29 of 41) and EAs (27 of 50).

Drug related variables were among the most significant positive predictors of opioid cessation in AAs and EAs including recency of last cocaine injection (OR_{AAs}=2.30 per level increase, $P_{AAs} = 9.11\times10^{-6}$) or use ($OR_{EAs}$=1.91 per level increase, $P_{EAs}=3.30\times10^{-15}$), while more years using heroin (OR_{AAs}=0.55 per standard deviation (SD) increase, $P_{AAs}=5.78\times10^{-6}$) or being older at first heavy opioid use (OR_{EAs}=0.56 per SD increase in age, $P_{EAs}=2.67\times10^{-12}$) decreased the odds of the outcome. Other drug use variables predicted greater odds of cessation but were less significant, such as “More time since last had alcohol symptoms lasting >1 month” (OR_{AAs}=1.45 per level increase, $P_{AAs}=2.84\times10^{-3}$) ($OR_{EAs}$=1.34 per level increase, $P_{EAs}=4.13\times10^{-5}$), “had 2 marijuana symptoms lasting >1 month” (OR_{AAs}=2.13, $P_{AAs}=4.83\times10^{-3}$) or “marijuana interfered with work or home activities” (OR_{EA}=1.67, $P_{EA}=1.61\times10^{-3}$), “smoked less frequently after waking up” (OR_{AAs}=1.75, $P_{AAs}=7.76\times10^{-3}$) or the Fagerstrom Test for Nicotine Dependence (FTND) item “able to cut down smoking” (OR_{EAs}=1.28, $P_{EAs}=3.69\times10^{-2}$). Having attended a self-help group for OUD (OR_{AAs}=1.72, $P_{AAs}=1.41\times10^{-2}$) or started attendance at an OUD self-help group sooner (OR_{EAs}=1.28 per level decrease, $P_{EAs}=2.4\times10^{-3}$) also increased the odds of cessation.

Several variables related to other mental health issues were also associated with opioid cessation. Self-harm (OR_{AAs}=1.39, $P_{AAs}=1.96\times10^{-2}$) or suicidal ideation ($OR_{EAs}$=1.2, $P_{EAs}=8.08\times10^{-3}$) were associated with significantly higher odds of cessation. Prior history of a depressive episode lasting >1 week

(OR_{AAs}=1.31, P_{AAs}=1.66×10^{-2}) or having drug-use associated depression (OR_{EAs}=1.64, P_{EAs}=2.55×10^{-3}) predicted higher odds of cessation. Pathological gambling severity (OR_{AAs}=0.8, P_{AAs}=3.32×10^{-2}) and no anxiety for longer than six month (OR_{AAs}=1.72, P_{AAs}= 2.06×10^{-3}) were significantly predictors of cessation in AAs. In EAs, recovering from an event causing PTSD assessed by the question "No fear in most disturbing/traumatizing event" (OR_{EAs}=1.93, P_{EAs}=1.66×10^{-6}), less recent antisocial behavior episodes (OR_{EAs}=1.35 per SD increase in time, P_{EAs}=1.03×10^{-4}), and unsafely raced cars (OR_{EAs}=1.78, P_{EAs}=3.79×10^{-3}) were associated with increased odds of cessation.

In AAs, female sex (OR_{AAs}=1.91, P_{AAs}=1.83×10^{-3}) and fulltime employment (OR_{AAs}=1.84, P_{AAs}=1.82×10^{-2}) were associated with a greater likelihood of opioid cessation, while having been raised primarily by a single parent (OR_{AAs}=0.63, P_{AAs}=1.3×10^{-2}) was associated with not achieving cessation. Other significant predictors of opioid cessation in AAs included HIV positive status (OR_{AAs}=2.47, P_{AAs}=1.39×10^{-3}), while in EAs, higher body mass index (OR_{EAs}=1.32 per SD change, P_{EAs}=2.58×10^{-6}), having asthma (OR_{EAs}=0.68, P_{EAs}=1.22×10^{-2}), higher household income (OR_{EAs}=1.15, P_{EAs}=1.3×10^{-3}), and being an atheist (OR_{EAs}=1.45, P_{EAs}=1.34×10^{-2}) were significant predictors.

Table 3.2. Stepwise regression results using all significant variables from all machine learning methods and models.

(A) African Americans and (B) European Americans.

A) African Americans

	Variable	**OR**	**p-value**
Drug related	Time since 1st opioid treatment*	1.56	1.90E-04
	Older age at first opioid misuse$	1.4	2.45E-02
	Older age at first opioid symptoms$	0.46	2.23E-05
	Diarrhea after stopping opioid use	1.54	3.68E-02
	Attended opioid self-help group	1.72	1.41E-02
	Older age at first heavy use of opioids$	0.73	2.70E-02
	N years using heroin$	0.55	5.78E-06
	Depressed after reducing cocaine use	0.53	4.05E-03
	Time since last injected cocaine*	2.30	9.11E-06
	Time since last used cocaine*	1.82	9.19E-05
	Time since last stayed high in cocaine*	1.41	2.93E-03
	Used cocaine <11 times within year of interview	2.67	1.38E-03
	Treated in outpatient program for cocaine use	1.88	4.06E-03
	Time since of first cocaine craving*	0.71	1.59E-03
	Never injected cocaine	2.53	1.75E-03
	Often used marijuana more than intended to	0.40	5.67E-04
	Mixed alcohol and drugs >3 times in 12 months	0.51	2.08E-03
	Time since last had alcohol symptoms lasting >1 month*	1.45	2.84E-03
	Drinking interfered with work/responsibility	0.60	2.71E-02
	Time since last attended alcohol self-help group*	0.79	2.65E-02
	Being alcohol dependent	1.73	2.56E-02
	Smoked in dangerous situation >3 times	0.66	3.75E-02
	Smoked less frequently after waking up	1.75	7.76E-03
	Older age at first cigarette$	1.31	6.24E-03
	Gave up social activities because of smoking	1.82	1.73E-02

	Time since first had sedatives[*]	1.3	1.06E-02
	Had 2 marijuana symptoms lasting a month	2.13	4.83E-03
	Disclosed drug problems to professionals	1.53	2.00E-02
	Number of years using sedatives[$]	1.29	1.33E-02
Behavioral	Pathological gambling severity	0.80	3.32E-02
	Time since last hurt oneself on purpose[*]	1.39	1.96E-02
	Time since last had depression >1 week[*]	1.31	1.66E-02
	Ever treated with medication or ECT for depression	1.75	4.33E-02
	Heard delusional noises when awake	1.61	3.40E-02
	No anxiety for >6 months	1.72	2.06E-02
Other Health	HIV positive	2.47	1.39E-03
	Health has always been better than now	0.62	9.64E-03
Demographic	Female sex	1.91	1.83E-03
	Raised primarily by single parent	0.63	1.30E-02
	Current age	2.44	1.41E-12
	Fulltime employment	1.84	1.82E-02

B) European Americans

	Variable	**OR**	**p-value**
Drug related	More time since last cocaine use*	1.91	3.30E-15
	Older age at first heavy opioid use$	0.56	2.67E-12
	Number of years using heroin$	0.69	3.01E-07
	More time since last cocaine injection*	1.85	2.38E-06
	More time since last had alcohol symptoms that last >1 month*	1.34	4.13E-05
	>20 outpatient visits for drug/psychiatric problems in the last year	1.76	9.56E-06
	More time since opioid treatment initiation*	1.53	2.24E-06
	Used cocaine >11 times in last year	0.47	1.09E-05
	More time since first used opioid 1/week for >1 month*	1.41	1.10E-04
	Have injected cocaine	2.01	1.66E-04
	Older age at first heavy cocaine use$	1.27	7.40E-04
	Marijuana interfered with work/home	1.67	1.61E-03
	More time since one started opioid self-help group*	1.28	2.40E-03
	More time since last feel high on cocaine for >1 day*	1.33	1.15E-03
	More time since last attended cocaine self-help group*	0.76	3.27E-04
	Continued using stimulant for its psychological problems	0.54	2.27E-02
	Heart slowed down when cutting down tobacco use	0.56	1.52E-02
	Used tobacco but not addicted	0.60	3.72E-03
	Disclosed problems with cocaine usage to professional	1.65	2.45E-03
	Always able to cut down smoking	1.28	3.69E-02

	Treated at outpatient drug program for cocaine	1.41	2.81E-02
	Used opioid more than intended to	0.63	1.15E-02
	Craved cocaine when cut down	1.42	1.05E-02
	Maximum N drinks in 24 hours[$]	1.15	1.61E-02
	Ever smoked >1 pack of cigarette daily for >1 month	1.16	2.16E-02
	Stopped using stimulants for >3 month	1.99	2.38E-03
	Drinking resulted in objections or problems with family and work	1.48	4.39E-03
Behavioral	Answered questions before they were completed as a child	0.73	2.24E-02
	<3 ASP criteria in 12 month period	0.64	2.69E-03
	Often failed to pay debts	0.68	2.17E-03
	Suspended or expelled from school	0.67	2.16E-03
	Number times in jail[$]	0.85	1.05E-02
	More time since last had suicide idea[*]	1.20	8.08E-03
	Less recent since last had antisocial behaviors[$]	1.35	1.03E-04
	No fear of most disturbing/traumatizing event	1.93	1.66E-06
	Avoided scenes that reminded of traumatic event	1.88	7.88E-05
	Had OCD behaviors when depressed	0.49	3.23E-04
	Feeling distracted	1.56	8.15E-04
	Unsafely raced cars	0.56	3.79E-03
	Found customers for prostitutes	0.68	4.04E-02
	Depression always started with drug problems	1.64	2.55E-03
	Had 8 depression symptoms	1.46	8.99E-03
	More time since last had depression episode[*]	1.20	1.65E-02
	Have outstanding emotional problem	1.63	5.55E-03

Other Health	Have asthma	0.68	1.22E-02
	Older age at heaviest weight[$]	1.21	2.68E-02
	Body mass index	1.32	3.59E-06
Demographics	Being an atheist	1.45	1.34E-02
	Household income	1.15	1.30E-03
	Current age	2.00	5.74E-09

OR=odds ratio
The values listed for each variable have been adjusted to take other variables into account.
* Categorical variables. 1- “within the last two weeks”, 2- “two weeks to just under one month ago”, 3-“one month to just under six months ago”, 4-“six months to a year ago”, 5-“more than a year ago”. OR value indicates the impact of per level change.
$ Continuous variables. OR value indicates the impact of per standard deviation change.

Discussion

We employed both regression and non-regression-based machine learning approaches to evaluate the association of more than 3,000 variables related to SUDs and other psychiatric disorders, other health-related behaviors, and demographic variables with opioid cessation among EAs and AAs assessed in a cross-sectional study of opioid, cocaine, and/or alcohol dependence. We observed moderate-to-high predictive accuracy across all methods; SVM, on average, marginally outperformed the other methods. Although the specific set of predictive variables differed in EAs and AAs, a common profile emerged. People who ceased opioid use tended to be older, initiated drug use later in life, had used opioids for a shorter period, experienced fewer problems related to cocaine or alcohol use, were currently employed, and had recovered from other psychiatric disorders including depression and PTSD compared to those whose opioid use persisted.

Previous research using machine learning for addiction outcomes focused mainly on predictive accuracy, although a few studies attempted to identify and interpret specific variables that were associated with the outcomes [163,164,198,199]. Acion et al. reported that ensemble super learning was superior to other machine learning methods, and used penalized regression, SVM, and neural networks for predicting SUD treatment success indicated by treatment discharge status in a Hispanic cohort [163]. In that study, less than 10% of participants had problems with cocaine or illicit opioids and fewer than 30 potential predictors were

assessed. In contrast, we evaluated several thousand predictors, including detailed measures of drug-use activities and psychiatric disorders, and ranked the importance of the top-ranked variables with four distinct machine learning algorithms. Gowin et al. identified regional brain activity changes predicting relapse from imaging data on fewer than 70 methamphetamine-dependent patients without including any lifestyle factors [163]. Che et al. applied deep learning to electronic health record data to identify people with short-term or long-term opioid use or dependence [164]. Similar to our study, they identified comorbid substance use and anxiety disorders as predictors [164]. Several other studies used only regression-based methods to predict opioid and stimulant dependence [200], cocaine dependence [201] and alcohol dependence [202], which might not capture other relationships among variables. Several of the non-regression-based methods we employed have also been applied in other studies, which focused mainly on MRI brain images as predictors of substance use disorder diagnoses [184,185,187,198].

We identified several predictors of opioid cessation that were previously associated with OUD or OUD-related conditions including co-morbid drug use, antisocial behavior, suicidal thoughts, HIV infection, and asthma [203-208]. Our finding that the majority of people who ceased opioids (60% in AAs and 66% in EAs) also ceased cocaine use is consistent with evidence of high rates of co-occurring OUD and cocaine use disorder (CUD) [209,210]. This finding also supports the use of treatment strategies that target both disorders [209,211] and suggests that

ceasing use of one substance might influence the ability to cease use of the other. Alternatively, ceasing both opioid and cocaine use may reflect self-selection for inclusion in the genetic studies in which 43% of AA and 32% of EA participants included those ascertained for CUD. Our findings are also consistent with observations that a failure to address tobacco use lowers the efficacy of opioid cessation treatment [212] and that a behavioral intervention in patients with antisocial personality disorder reduces substance use [213]. Unlike problems that are associated with other drug use, which predicts lower odds of opioid cessation, we found that cannabis use-related problems (e.g., two marijuana symptoms lasting a month, marijuana interfering with work) predicts higher odds of opioid cessation. This finding is puzzling and not immediately explainable. Previous findings of the co-occurrence of drug addiction, suicide attempts, depression, family conflicts, and PTSD, which may suggest bi-directional casual relationships [203-208], are consistent with our observation that better management of comorbid psychiatric problems (fewer recent suicide attempts and psychiatric symptoms) increases the likelihood of opioid cessation or vice versa.

We and Acion et al. identified age, employment status, and age at first drug use as factors for treatment success [163]. The protective effect of older age may be due to ascertainment bias because persons who survived severe dependence are more likely to have stopped using opioids. Full-time employment likely reduces the time or urge for persons dependent on opioids to seek and use the drug. In addition, drug screening associated with some jobs may reduce the

likelihood of current opioid use [214]. Quitting opioids also make it easier to find/maintain a job. The inverse correlation of age at first drug use and opioid cessation may reflect the increased difficulty of reversing the effect of long-term opioid exposure on the brain reward system [29] or increased severity associated with earlier onset.

Several significant predictors of opioid cessation related to non-substance-related behavior were population specific. Although these findings may be due in part to differences between AAs and EAs in willingness to endorse these behaviors, previous studies showed that AAs were more likely than EAs to report prolonged gambling and problems associated with gambling [215,216]. One explanation for our findings of significant associations of ceasing opioids with a self-reported HIV diagnosis in AAs is that OUD patients with severe or life-threatening illnesses are more likely to seek or adhere to treatment [217], an idea supported by evidence that HIV-infected patients have better treatment outcomes for OUD [218-220]. Alternatively, poorer general health may lead to reduced drug use [221] (the so-called “sick quitter”). In contrast, antisocial behavior, recovery from PTSD, and being an atheist were predictive of opioid cessation in EAs only. Prior research may provide insight into these EA-specific patterns. One study reported antisocial behaviors in EA children were significantly associated with substance initiation while the association was less strong in AA children [222], although the impact on opioid use was not assessed. PTSD and being an atheist identified in EAs might be due to the racial difference in exposure to traumatic events and

belief diversities [223,224]. For atheism, previous evidence about the effect of religion on SUDs is contradictory. One study showed that loss of religiosity between childhood and adulthood was associated with increased substance use while recent religiosity increased the odds of illicit drug use in the past year [225]. Alternatively, we could have observed similar factors in AAs but the sample size in AAs might limit our ability in identifying such associations.

The current study has several strengths. First, because the input dataset contains thousands of variables related to drug use activities, psychiatric disorders, medical history, and demographics obtained from several thousand individuals meeting DSM-5 criteria for OUD, we were able to explore many factors in addition to those included in other studies. Second, both linear and non-linear machine learning methods were employed to model the true underlying relationship between the predictors and outcome, which increased the number of factors we identified. Third, we evaluated three models for each machine learning method in order to better understand the contribution of opioid and other drug use information. Fourth, we considered only independent predictors in the association analyses to prevent over-representation of correlated factors. Finally, although there is no published “gold standard” predictive model against which to compare our results, the 80% predictive accuracy we achieved is similar to that seen in other machine learning studies [163,164,198,199].

Limitations of this work should be noted. First, given the cross-sectional

nature of our data and the over 90% relapse rate for OUD [226], many individuals classified as not using opioids may have subsequently relapsed to opioid use. However, it has been shown that prior abstinence is predictive of future abstinence, therefore people who ceased opioids are more likely to cease again even when relapse happened [227]. Second, the machine learning analyses were based on samples that may have been underpowered to detect associations with some variables compared to other studies that included tens of thousands of individuals [228]. Third, most persons in our cohort were evaluated prior to the current opioid epidemic and may not reflect recent secular trends in the prevalence and associated features of OUD. Fourth, associations of cessation with some variables and overall prediction accuracy may have been inflated because our analysis did not fully account for familial correlations. Fifth, in spite of the large number of variables that were included in the machine learning analyses, potentially important variables such as the reasons for first use and details of treatment and support programs were unavailable. However, we identified attending an opioid addiction self-help group as predictive of successful cessation, which is consistent with the reported benefit of self-help groups [229]. Sixth, the rate of response to many interview questions was substantially higher in EAs, while the sample size was twice that of AAs, which could account for some of the observed racial differences in predictive models. Given these limitations, our findings require external validation in larger samples before they can be incorporated in prediction models for clinical purposes. Finally, while

some of the factors identified were plausible and consistent with prior studies, other factor such as atheism is not immediately interpretable, such as “Found customers for prostitutes”. Given that our research is atheoretical, results should be interpreted with caution and be validated before implemented in clinical practice.

Conclusions

Using machine learning techniques with feature selection, we analyzed a large number of variables include demographic, behavioral, health and drug use activities and found variables in a wide range of domains that were associated with cessation. These included some that are consistent with prior literature, plausibly could be associated but have not been well studied, and do not have readily apparent explanations for their associations. Our findings suggest hypotheses for future studies and could inform how one might increase the likelihood of cessation with and without treatment. These results also support several widely known treatment strategies for OUD, such as treating psychiatric comorbidity, adding wraparound services like employment counseling, and simultaneously addressing polydrug use problems. Finally, in an era of increasing availability of digitized health-related records, our study provides a framework for disease outcome prediction using high dimensional phenotypic data purpose-collected via a research instrument.

.

CHAPTER FOUR: Cross-species Identification of Gene Networks and Enriched Pathways for Mouse Opioid Behaviors and Human Opioid Cessation

Abstract

Mice and human share about 99% genetic gene homologs. The use of mice as a model organism permits a comprehensive approach with precise environmental controls for identifying the effects of genetic influence on addiction behaviors, including drug sensitivity and spontaneous withdrawal. While there is a growing literature on genetic factors underlying various addiction traits have been reported in mice, a limited number of studies have cross-referenced human findings in mice. This translational approach is especially relevant because it provides additional leverage to protect against the potential high false positive rate of human genome-wide association studies (GWAS) findings. Therefore in this study, we conducted weighted gene co-expression network analysis (WGCNA) on a reduced genetic complexity cross between mouse substrains that displaying differential opioid sensitivity and withdrawal behaviors owing to an underlying QTL on distal chromosome 1. We cross referenced trans-regulated gene homologs and pathways originating from this locus that were also correlated with mouse behavior and identified gene modules that could then be compared to opioid cessation associated genes from human GWAS. Among the top overlapping genes, two genes *HNRNPU* and *GRB10* were identified to have translational addiction-relevance. Pathway analysis identified several enriched

gene sets that were previously known to be regulated by opioids, as well as novel pathways that provide potential new mechanistic insights to the shared genetics between mice and human in response to chronic opioid use

Introduction

The genetics of opioid addiction in human and animal models are heavily studied[106,230-232]. We identified variants that were associated with opioid cessation in people with opioid use disorder (OUD) through genome-wide association studies (GWAS). The genes of the top variants (Chapter Two) are highly expressed in brain and have relevance to addiction phenotypes and psychiatric traits. However, false positive genetic association results have always been a concern in GWAS interpretation[233], which makes animal models attractive for exploring the extent of concordance between these two lines of evidence[234].

Mouse models are an extremely useful tool for identifying the genetic factors contributing to addiction-relevant phenotypes[235]. The large overlap between mice and human gene homologs (99%)[236] allows for cross-species translation to human[73]. Furthermore, the environmental and genetic factors influencing addiction behaviors in mice can be precisely altered and the tissue specific gene expression level in response to genetic and environmental alterations can be closely monitored with high-throughput sequencing techniques. These attractive features of mouse models overcome the ethical limitations in using human subjects and make it possible to study clinically relevant addiction traits in a refined manner. While many methods used

experimenter-induced abstinence and relapse[237], measuring voluntary opioid cessation in mice remains to be challenging. Hyper-locomotion induced by drug abuse is a heritable behavior that is linked to stimulation of the mesolimbic reward pathway[92] and anxiety-like behavior is a hallmark of opioid withdrawal[238]. , Genetic analysis of these traits in mice might allow one to cross-reference and translate findings to human opioid cessation.

Weighted gene co-expression network analysis (WGCNA)[239] evaluates the co-expression patterns of genes based on their connectivity using correlation methods. WGCNA permits integration of transcriptome data with both genetic and phenotypic data (e.g., behavior) to identify gene networks that are trait-relevant and has been widely used in both human and mouse studies[240,241]. Tools that allow for cross-species comparisons are trending. GeneWeaver[242] is a web based tool that integrates experimental results with published GWAS, quantitative trait locus, microarray, RNA-sequencing, and mutant phenotyping studies to identify gene-function associations across diverse experiments, species, conditions, behaviors or biological processes. GeneWeaver includes functions such as homology integration tool that allows translation across species and similarity measurement between gene sets for one to assess the degree of overlap.

The goal of the current study is to use WGCNA to refine the trans-regulated gene list from mouse transcriptomic findings in response to opioid induced behaviors, so that more accurate results can be obtained when

comparing with genes identified from human GWAS. A previous study of oxycodone (OXY) induced behavioral quantitative trait locus (QTL) mapping on F2 reduced complexity crosses derived from B6J and B6NJ F1 intercross identified distal chromosome (chr) 1 (163-190 Mb) as a solely responsible locus for opioid sensitivity and opioid withdrawal behaviors[243]. Previously, expression QTL analysis identified extensive trans-regulation of gene expression associated with genetic markers distinguishing C57BL/6 substrains[243]. A gene list with 3495 genes was obtained by applying 0.5% FDR cutoff to control for false positive on trans-regulated genes[244]. We used this gene list in WGCNA to identify gene networks (modules) that were correlated with opioid addiction behaviors and conducted a cross species comparison with the gene list for human opioid cessation genes that top variants are in. Our findings provide a framework for cross-species reference between humans and mice in addiction genetics.

Material and methods

Mice

The founder mice C57BL/6J (B6J) and C57BL/NJ (B6NJ) mice were purchased from The Jackson Laboratory (Bar Harbor, ME) at 7 weeks of age. B6J females were crossed to B6NJ males to generate B6J x B6NJ-F_1 mice and then B6J x B6NJ F_1 offspring were intercrossed to generate B6J x B6NJ F_2 mice. Mice within the same cage receive the same treatment. F_2 mice were selected based on genotypes at distal chr1 (from the SNP array; 163 Mb, 181 Mb) to backcross to B6J to generate mice that were either homozygous B6J (J/J) or

heterozygous (J/N) across distal chr1. Other variations on the genome were considered unlikely to affect the Mendelian trait of Day 2 OXY-induced locomotor activity, therefore we focused solely on recombination events at distal chr 1 (163-190 Mb). 11 mice (4 females) with J/J genotype at chr 1 marker and 12 mice (3 females) with J/N genotype at chr1 marker were tested.

Previous Behavioral Testing, Tissue Collection, Transcriptome Analysis and trans-eQTL analysis

Appendix II is a detailed description of the methods. Behavioral studies including OXY conditioned place preference, OXY induced locomotor activity (measured in Day 2 and Day 4), and elevated plus maze were done by Lisa Goldberg[243]. The striatum were collected for mRNA sequencing 24 hours after elevated plus maze testing from 23 OXY-treated F_2, who were either homozygous (J/J) or heterozygous (J/N) at the markers capturing a QTL on distal a chr 1 for OXY-induced locomotor activity (rs6341208-163.13 Mb, rs51237371-181.32 Mb). Transcriptomic analyses were done by regressing VOOM[245] normalized gene expression levels to the chr1 marker, adjusting for sex. Trans-eQTL analysis were done by regressing gene expression levels to 90 SNPs throughout the genome using a linear model with an additive model in R package Matrix eQTL[246].

Weighted Gene Co-expression Network Analysis (WGCNA)

The differentially expressed genes from above were annotated by R BioMart[247] package using *musculus* Ensembl database (www.ensembl.org). VOOM transformed read counts averaged over sample replicates on genes that were differentially expressed in response to chr 1 marker were used as the input for WGCNA[239]. Using WGCNA, we identified no sample outliers, where hierarchical clustering dendrogram was performed on the similarity measurement of normalized read count. Using 1-step network construction and module detection, correlated gene clusters were identified. In brief, Pearson correlation between genes were calculated to capture the linear-relationships between genes. A soft thresholding power of 10 was chosen based on the criterion of maximizing approximate scale-free topology (scale free topology model fit =0.9) and applied on the Pearson correlation matrix to get the “scale-free” adjacency matrix. The adjacency matrix was transformed to topological overlap measures (TOM) to account for neighborhood similarities. The co-expression modules were detected by hierarchical clustering from TOM, which yielded 14 gene clusters with between 61 and 1140 genes. Gene clusters (modules) were given different color names, with “grey” indicating genes that don’t belong to any specific modules. We further related behavioral traits and genotypes observed to these gene clusters (modules) for association, where the first principal component (module eigengene) of a given module was individually correlated to the traits mentioned above. Genes within modules identified with eigengene significance

($p<0.001$) were retrieved and plotted against their significant level of a given trait defined by the absolute value of the correlation between gene and the trait. Module membership was defined as the correlation of the module eigengene and the gene expression profile.

Gene Enrichment Analysis

Genes within modules whose eigengenes were significantly correlated ($p<0.001$) with a chr 1 marker were used and combined with human cessation GWAS top genes as input to Enrichr[248], an online enrichment analysis tool with databases including Gene Ontology (GO) Biological Process, GO Cellular Component, GO Molecular Function, and KEGG. Enrichr uses Fisher's exact test to calculate an enrichment score, the test for each pathway was computed by evaluating its observed rank and how much it deviates from expected rank using several random input gene lists. The final output includes p value from Fisher's test and Z statistics from rank test as well as a combined score, which is calculated as $\log(p)*Z$.

Cross-species Reference Using GeneWeaver

To find the similarities between human GWAS genes to thousands of mice differentially expressed genes (DEGs), we used a less stringent cutoff ($p<0.05$, 1765 genes) on human GWAS variant mapped genes whose p values were corrected by Li and Ji method[249] for the effective number of independent SNP tests performed within each gene. Genes within trait-significant modules from

mice trans-eQTL genes were used as mouse input. Homologs between human and mouse gene lists mentioned above were identified with two types of comparisons – mouse significant module genes vs. human opioid cessation gene list stratified by modules as well as mouse all trans-eQTL gene lists vs. the human opioid cessation gene list - using “GeneSet Graph” tool from GeneWeaver (www.geneweaver.org)[242] . “GeneSet Graph” allows one to visualize the intersections of genes from multiple datasets, where homologs of different forms of gene and species can be identified. The same two types of comparisons between two genes sets mentioned above were performed to find the similarity between two gene sets via “Jaccard similarity”[250] tool from GeneWeaver, where the similarity coefficients were measured as the size of the intersection of two gene sets divided by the size of the union between two gene sets, its associated p values were returned as well.

Results

Chr1 marker and OXY-induced traits associated gene clusters

We identified 15 modules (co-expressed gene clusters) with the size ranging from 61 to 1140 genes among 3495 trans-regulated genes (Figure 4.1). Pink and purple module eigengenes were significantly correlated with chr1 marker rs51237371 genotype (r_{pink}=-0.87, P_{pink} = 5×10^{-8}; r_{purple}=0.62, P_{pink} = 1×10^{-3}). Purple module eigengene was more significantly correlated with OXY-induced locomotor activities and spontaneous withdrawal behavior than pink module ($r_{pink\_d2\_distance}$=0.51, $P_{pink\ d2\_distance}$ = 0.01; $r_{purple\_d2\_distance}$=-0.51, $P_{purple\ d2\_distance}$ =

0.01; $r_{pink\_d4\_distance}$=0.34, $P_{pink\ d4\_distance}$ = 0.1; $r_{purple\_d4\_distance}$=-0.55, $P_{purple\ d4\_distance}$ = 6×10^{-3}; $r_{pink\_prctopenarm}$=0.36, $P_{pink\ prctopenarm}$ = 0.09; $r_{purple\_prctopenarm}$=-0.42, $P_{purple\ prctopenarm}$ = 0.05). Although the degree of the correlations were similar between pink and purple module eigengenes to genotype and behavioral traits, pink module eigengene was negatively correlated with chr 1 genotype and positively correlated with behavioral traits, while the opposite was true for purple. Sex was not correlated with any module.

Genes in both pink and purple modules showed significant correlation to behavioral traits (Figure 4.2). Genes in the pink module had stronger correlation to locomotor activity and OXY induced withdrawal ($r_{pink\_d2\_distance}$=0.52, $P_{pink\ d2\_distance}$ = 4.9×10^{-9}; $r_{purple\_d2\_distance}$= 0.35, $P_{purple\ d2\_distance}$ = 2.4×10^{-3} ; $r_{pink\_prctopenarm}$ =0.35, $P_{pink\ prctopenarm}$ = 1.7×10^{-4}; $r_{purple\_prctopenarm}$ = 0.41, $P_{purple\ prctopenarm}$ = 3.2×10^{-4}). In addition, genes in the pink module were significantly correlated with chr1 marker genotype, purple module genes were not correlated with chr 1 marker genotype ($r_{pink\_chr1\_marker}$ =0.67, $P_{pink\_chr1\_marker}$ = 8.9×10^{-16}; $r_{purple\_chr1\_marker}$ = 0.19, $P_{purple\_chr1\_marker}$ = 0.11).

Figure 4.1 Module eigengene – opioid trait relationship among trans-regulated genes in mice.

Co-expressed gene clusters (modules) were shown with prefix “ME” and a name of a color. “sex” :sex of the mice, “rs51237371_1”: genotype of chr1 marker, which can be either J/J or J/N. “D2.Distance”: day 2 (first OXY exposure) locomotor activity between chr 1 marker J/J and J/N mice measured in meters. “D4.Distance”: day 4 (second OXY exposure) locomotor activity between chr 1 marker J/J and J/N mice measured in meters. “prctopenarm”: percent of time spent in the open arm of elevated plus maze. For each cell, the level of correlation is shown followed by the p value in parentheses.

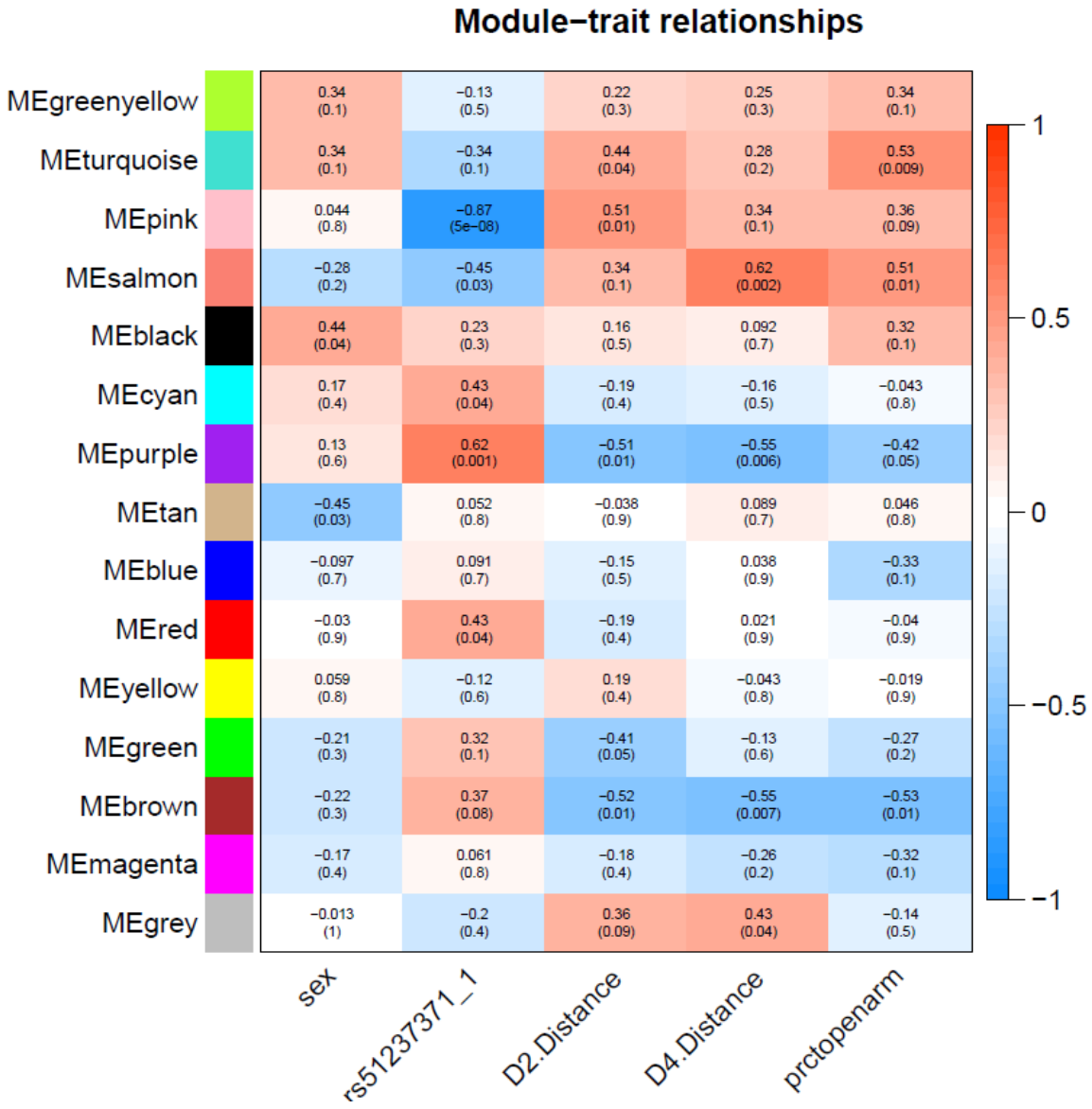
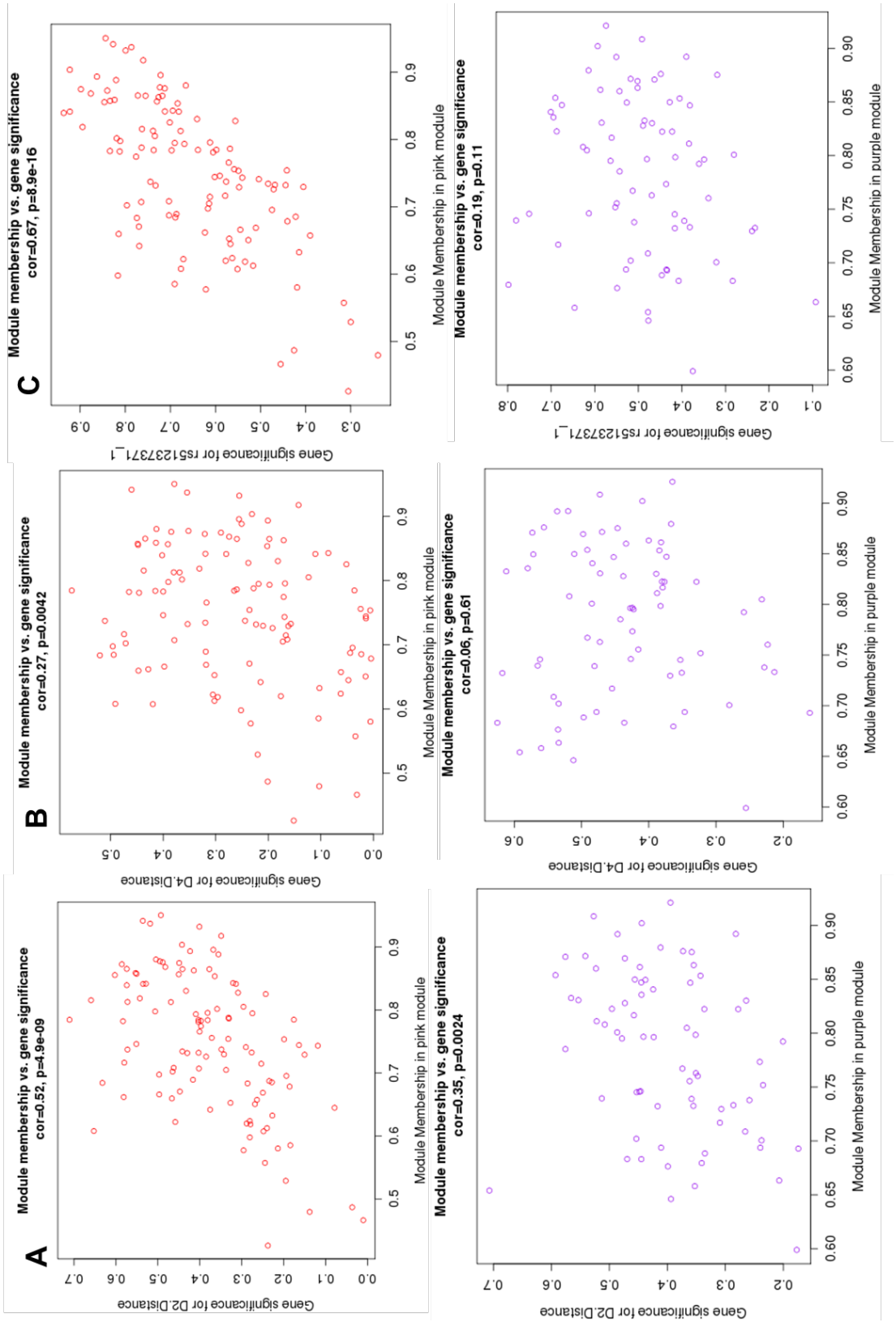

Module−trait relationships
MEgreenyellow
MEturquoise
MEpink
MEsalmon
MEblack
MEcyan
MEpurple
MEtan
MEblue
MEred
MEyellow
MEgreen
MEbrown
MEmagenta
MEgrey
0.34 (0.1)
−0.13 (0.5)
0.22 (0.3)
0.25 (0.3)
0.34 (0.1)
0.34 (0.1)
−0.34 (0.1)
0.44 (0.04)
0.28 (0.2)
0.53 (0.009)
0.044 (0.8)
−0.87 (5e−08)
0.51 (0.01)
0.34 (0.1)
0.36 (0.09)
−0.28 (0.2)
−0.45 (0.03)
0.34 (0.1)
0.62 (0.002)
0.51 (0.01)
0.44 (0.04)
0.23 (0.3)
0.16 (0.5)
0.092 (0.7)
0.32 (0.1)
0.17 (0.4)
0.43 (0.04)
−0.19 (0.4)
−0.16 (0.5)
−0.043 (0.8)
0.13 (0.6)
0.62 (0.001)
−0.51 (0.01)
−0.55 (0.006)
−0.42 (0.05)
−0.45 (0.03)
0.052 (0.8)
−0.038 (0.9)
0.089 (0.7)
0.046 (0.8)
−0.097 (0.7)
0.091 (0.7)
−0.15 (0.5)
0.038 (0.9)
−0.33 (0.1)
−0.03 (0.9)
0.43 (0.04)
−0.19 (0.4)
0.021 (0.9)
−0.04 (0.9)
0.059 (0.8)
−0.12 (0.6)
0.19 (0.4)
−0.043 (0.8)
−0.019 (0.9)
−0.21 (0.3)
0.32 (0.1)
−0.41 (0.05)
−0.13 (0.6)
−0.27 (0.2)
−0.22 (0.3)
0.37 (0.08)
−0.52 (0.01)
−0.55 (0.007)
−0.53 (0.01)
−0.17 (0.4)
0.061 (0.8)
−0.18 (0.4)
−0.26 (0.2)
−0.32 (0.1)
−0.013 (1)
−0.2 (0.4)
0.36 (0.09)
0.43 (0.04)
−0.14 (0.5)
sex
rs51237371_1
D2.Distance
D4.Distance
prctopenarm
1
0.5
0
−0.5
−1

Figure 4.2 Scatter plots of gene significance against module membership of OXY-induced locomotor activity.

The locomotor activity of the first OXY exposure (A), OXY-induced locomotor activity of the second OXY exposure (B), OXY-induced withdrawal (C), and chr1 marker genotype (D) using genes in pink and purple modules. Gene significance is defined as the absolute value of the correlation between the gene and the trait. Module membership is defined as the correlation of the module eigengene and the gene expression profile. “rs51237371_1” indicates genotype of chr1 marker, which can be either J/J or J/N. “D2.Distance” indicates day 2 (first OXY exposure) locomotor activity between chr 1 marker J/J and J/N mice measured in meters. “D4.Distance” indicates day 4 (second OXY exposure) locomotor activity between chr 1 marker J/J and J/N mice measured in meters. “prctopenarm” indicates percent of time mice stayed at the open arm of elevated plus maze.


A
Module membership vs. gene significance
cor=0.52, p=4.9e-09
Gene significance for D2.Distance
Module Membership in pink module
Module membership vs. gene significance
cor=0.35, p=0.0024
Gene significance for D2.Distance
Module Membership in purple module
B
Module membership vs. gene significance
cor=0.27, p=0.0042
Gene significance for D4.Distance
Module Membership in pink module
Module membership vs. gene significance
cor=0.06, p=0.61
Gene significance for D4.Distance
Module Membership in purple module
C
Module membership vs. gene significance
cor=0.67, p=8.9e-16
Gene significance for rs51237371_1
Module Membership in pink module
Module membership vs. gene significance
cor=0.19, p=0.11
Gene significance for rs51237371_1
Module Membership in purple module

Gene Set Enrichment of Chr1 Marker Associated Modules

One pathway was significantly enriched ($P_{adj}<0.05$) using genes in the purple module whereas six pathways were significantly enriched using genes in the pink module (Table 4.1). No pathways with the same names overlapped with human enriched pathways. Combining human GWAS finding genes with purple or pink module for pathway enrichment tests did not result in any pathway that overlaps with the human findings.

Genes that Overlap between Chr1 Marker Associated Modules in Mice and Human Opioid cessation

Among 1756 human genes and 73 genes within the purple module, 4 genes *DLST, MFN2, HNRNPU* and *EPDR1* overlapped. Among 1756 human genes and 111 pink module genes, 5 genes *TIAM1, ARHGAP12, SP4, RSF1* and *GRB10* overlapped. Among 1756 human genes and 3495 trans-regulated mouse genes, 232 genes overlapped. The overlap observed for each pair was small but significant ($P_{pink\_module\_to\_human}<0.002$, Jaccard similarity (J) $_{pink\_module\_to\_human}=0.003$; $P_{purple\_module\_to\_human}<0.002$, $J_{purple\_module\_to\_human}=0.0038$; $P_{all\_mice\_trans\_eqtl\_to\_human}<0.002$, $J_{all\_mice\_trans\_eqtl\_to\_human}=0.0515$).

Table 4.1 Enriched gene set using chr 1 marker associated modules in mice with and without genes from human opioid cessation.

Purple, pink: module name. hubgene: the gene with the highest connectivity within a module. Representative genes: genes in the module that were also in the identified pathway. GWAS genes: 89 genes identified from human opioid cessation GWAS result adjusted for gene size and LD.

	N	hubgene	Enriched Pathways	P_{adj}	representative genes
purple	73	Fchsd2	reduced number of immature B cells	0.04	SPPL2A, ATP11C
pink	111	Tspyl4	Glycolysis and Gluconeogenesis	0.008	LDHA, GOT1, ENO1B, ALDOA
			Mitotic G2-G2/M phases	0.036	ACTR1A, HSP90AB1, FBXW11, PSME3, DYNLL1, BTRC
			mTOR signaling pathway	0.003	RRAGA, FBXW11, BNIP3, RAC1, EEF2
			Asparagine and aspartate biosynthesis	0.004	GOT1, ASNS
			ATP synthesis	0.004	ATP5B, ATP5A1
			Distal sensory impairment	0.036	ABHD12, YARS, MFN2, GARS
purple + GWAS genes	162		-		
pink+ GWAS genes	200		TRIB3 kinase	2.00E-05	YARS, ARF1, TMED9, GOT1, ASNS, EEF2, LURAPIL, BSG, MICALL1, EIF4H, GARS, CHAC1, EIF3C, ALDOA, ATF4, C11ORF24
			MAP3K7 kinase	7.00E-03	NARS, APIS, MORF4L2, UBE2D3, SERBP1, GDI12, G3BP2, EIF4H, CDK1, HNRNPU, GARS, MZT1
			CSNK2A kinase	7.00E-03	YARS, NARS, ATP5B, CXADR, HSP90AB1, APIS, SERBP1, GDII, G3BP2, HNRNPU, CHD4, CDC25A
			MAPK6 kinase	2.00E-02	NARS, ATP5B, HSP90AB1, APIS, SERBP1, DNAJA2, GDI2, G3BP2, ASNS, GARS
			MTOR	2.00E-02	KMT2D, SERBP1, EIF4H, MFN2, HNRNPU, DLST, ADAR, CHD4, EEF2, CDC25A

Discussion

Overlapping Addiction Risk Genes in Humans and Mice

Among genes that overlap between Chr1 marker associated modules in mice and humans, two genes, *HNRNPU* and *GRB10,* have previously been associated with addiction-related behaviors. *HNRNPU* is heterogeneous nuclear ribonucleoprotein U, which encodes a protein that bind both RNA and DNA, and involved in the formation of ribonucleoprotein complexes. *HNRNPU* was identified as a gene with the most connectivity (hubgene) in mouse alcohol dependence-associated gene networks sampled from the nucleus accumbens[251]. *HNRNPH*, a protein in the subfamily of *HNRNPU*, is known to interact with an intronic SNP[252,253], which leads to alternative splicing of the opioid receptor gene *OPRM1* in humans[254]. Another protein in this subfamily, *HNRNPH1,* has also been studied for its influence on methamphetamine response in mice[253,255]. The second identified overlapping gene *GRB10* encodes growth factor receptor bound protein 10. Mice lacking paternal *GRB10* make fewer impulsive choices, such impulsive choices could contributes to pathological conditions such as gambling and drug addiction[256]. Variants in *GRB10* were also identified for substance dependence in humans[257]. Other research on *GRB10* were mainly about its potential tumor suppressing and cell renewal effects[258-260].

Implication of Enriched Gene Set

No pathway terms overlapped between the mouse chr1 marker associated enriched gene sets and human opioid cessation GWAS enriched gene sets. However, pathways enriched in the human opioid cessation GWAS are known to interact with pathways identified from the mouse analyses. For example, the vitamin D metabolism pathway identified from the human opioid cessation GWAS has an effect in mTOR signaling[261] identified in mice enriched gene set, although the representative genes of these pathways did not overlap. Vitamin D, a key player in vitamin D metabolism pathway, also inhibits B cell proliferation and blocks B cell differentiation and immunoglobulin section[262], which might result in "a reduced number of immature B cells". As "a reduced number of immature B cells" was identified as an enriched gene set from mice. Loss of B cells in both human and in animal models has been observed after chronic exposure to alcohol and opioids[263,264]. Glycolysis and Gluconeogenesis is a enriched gene set from mice, and this gene set includes the process of cells converting glucose to lactate followed by the synthesis of ATP (glycolysis) and the process of glucose biosynthesis from non-carbohydrate precursors (gluconeogenesis)[265]. There is evidence of a link between chronic consumption of opioids and delayed insulin response to food ingestion, which leads to increased hepatic glucose production secondary to glycogenolysis[266,267]. Similarly, the FGF signaling pathway enriched in the human GWAS is also known to regulate energy metabolism[268]. However, whether the FGF signaling pathway and Glycolysis and

Gluconeogenesis pathway are functioning together in the presence of opioid remains to be studied.

Limitations

Several limitations should be noted. First, the mouse model used for this analysis was not designed to study opioid cessation but rather to study OXY-induced sensitivity and withdrawal. A study design using mice with chronic opioid exposure with and without long-term abstinence would better model the opioid exposure in human cessation. Second, although the environment factors can be precisely monitored and studied for addiction behaviors in mice, factors affecting opioid cessation that are unique to humans such as drug price and social-economic factors could alter the outcome, and are challenging to model in mice. Opioid abstinence in mice would likely to be achieved by controlling the drug accessibility. In this study, using F2 reduced complexity crosses as a mouse model further decreased genetic complexity, which affects the number of genes identified for associations with addiction phenotypes thus limiting the degree of cross-species gene overlap. Fourth, only the striatum was analyzed in mice, other brain regions such as prefrontal cortex, which regulates brain reward, self-control, and drug craving[269], were not studied. Fifth, human brain transcriptome data was not available for this project. Having human striatal RNA-seq data from the same opioid cessation participants would give us a more direct comparison. Finally, the gene set enrichment software used in mice was different from the one

used in human, thus could lead to less comparable results. Due to the limited power, we didn't find any enriched gene set using Enrichr in human and the enrichment software used in human from Chapter Two was using GWAS summary statistics as input, thus could not be applied directly to mice data.

Conclusion

Using trans-regulated genes identified from F2 reduced complexity crosses that were previously shown to have different OXY-induced sensitivity and withdrawal behaviors caused by one or more polymorphisms in linkage disequilibrium with a chr 1 marker, we constructed a biological network based on the co-expressed gene modules in striatum. We identified two gene modules that were associated with both the chr 1 marker genotype and behavioral traits. Cross-species gene homologs and pathways were identified between mouse OXY-related genes and human opioid cessation GWAS mapped genes. Our findings suggest novel genes and potential pathway mechanisms that are shared between mice and human in response to chronic opioid use.

CHAPTER FIVE Discussion and Future Directions

Summary

This dissertation investigated the genetic and non-genetic basis of opioid cessation defined by recency of last opioid use among people who meet at least mild lifetime OUD criteria. Additionally, we used a mouse model that displayed differential opioid withdrawal behaviors to find genetic overlap between mouse and human through gene co-expression network analysis and pathway analysis.

Research on opioid cessation is extremely sparse. This is likely due in part to high relapse rates and clinical ambiguity of cessation; therefore, the majority of studies have so far focused on treatment completion rates as the primary outcome[174-176]. However, this outcome has a selection bias whereby people who have received treatment are more likely to be included in the analysis. In contrast, using opioid cessation as the primary outcome allows inclusion of individuals regardless of their treatment status. These analyses, in theory, might have greater potential to detect factors related to opioid cessation that are independent from treatment status.

This thesis includes the first genetic study of opioid cessation (Chapter Two). Studies investigating the genetic basis of opioid use have typically focused on opioid dependence or OUD severity level[39,40] rather than opioid cessation. We utilized all currently available genetic datasets (to the best of our knowledge) to study this phenotype. Although our sample size was much larger than most studies investigating outcomes related to opioid cessation, it is relatively small for

a GWAS, resulting in insufficient power to detect genome-wide significant associations. Nonetheless, the top signals identified can serve as a reference for similar research.

Our association study of non-genetic factors for opioid cessation (Chapter Three) is unique in two ways. First, compared with other studies, the number of variables derived from the structured psychiatric interview is multiple orders of magnitude greater than most other studies of this outcome. Second, we used both non-linear and linear machine learning models to identify many possible combinations of variables that are predictive of cessation. In addition, the analyses were performed using factors that were grouped in multiple ways to identify contributions of opioid-related, other drug-related, and non-drug variables that significantly and independently predict opioid cessation. Using this information, we identified factors related to treatment success or other drug use problems that have been previously reported by others, as well as novel factors some of which might not be immediately explainable. Our proof-of concept study provides a framework for identifying disease related factors in healthcare data and also generates hypotheses for future studies.

Finally, we used a mouse model that displays varied opioid response and opioid withdrawal behaviors to identify gene co-expression patterns related to opioid behaviors (Chapter Four). While opioid cessation is difficult to model in animals, opioid response and opioid withdrawal severity are traits likely to be associated with relevant cessation outcomes in humans. Using RNA-seq

profiling, we identified several pathways enriched with specific genes clusters, and while these pathways do not overlap completely with human findings from Chapter Two, these pathways have been shown to interact with each other with respect to opioids. Furthermore, our comparison of the top genes associated with opioid cessation in humans and opioid behaviors in mice revealed several gene homologs that are related to impulsivity in humans, a behavioral trait known to be strongly associated with addiction.

Future Directions

The work presented in this thesis suggests several future directions for continued exploration of factors related to opioid cessation. First, an epidemiological study should recruit a larger number of individuals to study opioid cessation as a primary outcome. Analyses conducted in such a dataset would have higher statistical power than currently available datasets, and will also enable analyses for identifying robust associations with rare variants. In future genetic studies, it will be advantageous to collect both gene-expression and epigenetic data from the study sample in order to evaluate the casual relationship between genetics and the outcome of interest. Second, in large scale studies of drug addiction research, there should be a greater emphasis on collection of more in-depth information regarding drug related activity (e.g., treatment clinic attendance, treatment frequency, reason of first drug use) in order to increase precision and robustness of associations with genetic and non-genetic factors. Third, as computational models become more advanced, we should explore data

using more than one machine learning model on top of existing models, such as the hybrid methods (stack several machine learning models together). Fourth, increased focus of RNA-sequencing brain regions most relevant to addiction (e.g. nucleus accumbens) would allow development of tissue specific networks and creation of tissue-specific gene profiles for each withdrawal behavior. Finally, the framework employed in this thesis project for using both genetic and non-genetic data to identify disease relevant factors, as well as animal models to enhance human findings, could be applied to other diseases to provide genetic, molecular and behavioral insights for better disease management.

APPENDIX I

Supplemental Methods for Chapter Three

Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO implemented in the R package 'glmnet'[270] was used for both feature selection and prediction. The shrinkage parameter lambda in the penalty term of LASSO regression was obtained using 10-fold cross validation on the training set 10 times. Separate accuracy criteria of either misclassification error or AUC were used to search for the lambda with the best model fit. The "1SE rule[271]" which aims to find the simplest model with comparable accuracy to the best model, was used to identify lambda whose cross validation error was one standard error unit from the lowest cross validation error on the training set. We identified the significant features by fitting lambdas on the training set. The test set accuracy was evaluated by using the class probability prediction of test set as an input to Scikit-learn[197] (roc_auc_score and f1_score) to obtain our final accuracy measures: F1 score and AUC.

Support Vector Machine (SVM) with Recursive Feature Elimination

Linear SVM with recursive feature elimination was implemented using the Scikit-learn python package. A soft margin was included in the model to reduce overfitting. Briefly, we set the penalty parameter C by exhaustively searching the range of values between 2^{-20} and 2^{8} that covers the recommended range proposed by Hsu et.al[190] using 10-fold cross validation with balanced class

weights on the training set . We applied the parameter C chosen by either maximizing AUC or F1 score (whichever yielded the highest accuracy) in the training set for recursive feature selection. A feature's importance, represented by its weight, was used as the input to the recursive feature elimination function, where 10-fold cross validation was used to find the combination of features that maximizes either AUC or F1 score (whichever yielded the highest accuracy). Selected features from each model were applied to the test set to obtain the overall test accuracy. Given that RFE only evaluates a limited number of feature importance criteria, only the linear kernel was used for its ability to yield a smaller feature importance space, such that the feature importance of each variable is explained by one value instead of a collection of similarity values in reference to other variables.

Random Forest (RF) Recursive Feature Elimination

The RF recursive feature elimination approach was implemented using the Scikit-learn python package. We determined the ideal number of decision trees needed for the RF model by searching the recommended range of 2^7 to 2^{11} proposed by Oshiro et.al[191] 10-fold cross validation with balanced class weight on the training set. The decision trees that associated with the maximum AUC or maximum F1 score were individually chosen. A feature's importance, represented by the Gini impurity[272] measure proposed by Breiman[273], was used to evaluate the importance of a variable by adding up the weighted impurity decreases for all nodes averaged over all decision trees. A feature's importance

rank was used as the input for recursive feature elimination and 10-fold cross validation was used to find the combination of features that maximized either AUC or F1 score (whichever yielded the highest accuracy). Selected features from each model were included in the test set without any feature selection steps to obtain the test accuracy.

Deep Neural Network (DNN)

The DNN approach was applied using Keras[274] with a tensorflow framework. A 3-layer, fully connected feed forward DNN was constructed with two hidden layers using rectified linear units (Relu) as an activation function and a sigmoid function for the output layer. The training set that was used in the aforementioned methods was further split into a development set and validation set with 9:1 ratio with fixed case control ratios to minimize overfitting or inadequate training. The Adam optimization method was used to find efficiently the parameters associated with the ideal state of the DNN objective function. The Adam optimizer[275] has been proven to be the start of art[276-278] optimization method by incorporating the advantages of two most popular optimizations (RMSProp[279] and AdaGrad[280]). We followed the suggested hyperparameter settings from the developer of the Adam optimizer[275] and only tuned the learning rate. L2 normalization with various scales were performed on each layer to prevent overfitting. Cross validation of the development set and balanced accuracy were used to reduce bias for the hyperparameter search. Both weighted AUC and weighted F1 score on the validation set were used to

measure model performance.

Several methods have been proposed for feature selection using DNN, with the focus on reducing input dimensionality, such as sparse one-to-one, dropout feature ranking, and activation potential based[192,281,282]. We used the activation potential based method because of its proven performance in reducing the number of features, to obviate application of another filtering method coupled with DNN, and its simplicity and intuitiveness for selecting the number of important variables. Feature selection was performed according to the method proposed by Roy et.al[192]. Briefly, we computed the activation potential of each input feature connected to each of the hidden nodes in the first layer before applying Relu. Then, the average relative activation potential of each feature in each first hidden layer node was calculated by averaging the number of input features and training samples at each node. Relu activation was applied to the average relative activation potential to obtain the net positive contribution of each input feature. Input features were ranked and plotted against their net positive activation potential contribution. Important features were chosen based on their collective contribution of net positive activation potential. Because of bias associated with AUC in the presence of imbalanced dataset, the feature combination that associated with the highest F1 score in the validation set was used to obtain the accuracy on test set.

Performance of Machine Learning Algorithms

There was a consistent drop in accuracy from model 1 to model 3 across the four machine learning methods, although the difference between models 1 and 2 is smaller than the difference between models 2 and 3 (Figure S2). The loss of accuracy across models was greater in AAs than in EAs. The AUC generally demonstrated higher accuracy than the F1 score. Among the machine learning methods, SVM yielded the highest F1 score more frequently than the other methods across models in both AAs and EAs. Most notably, SVM had the best performance for model 2 in AAs and models 2 and 3 in EAs, although the differences in accuracy between SVM and the other models with high performance were small. The observation that SVM performed only marginally better than LASSO in both AAs and EAs was not surprising because SVM using a linear kernel and LASSO employ a linear model with regularization. Both SVM and LASSO selected uncorrelated features, however SVM also selected correlated features. Both RF and DNN showed some evidence for overfitting (result not shown), although the effect was relatively small that was reflected by an approximately 4% higher accuracy in the cross validation training set than the test set. The RF model may have been overfitted because the number of individual classifier decision trees was fixed. DNN generally requires a much larger sample size than the one available here, which might have limited its performance.

Appendix I Table S1. Five most significant variables ranked by feature importance for each machine learning method, stratified by model and population**.**

African Americans

		Variable	Rank	p-value
Model 1	LASSO	Recency of last cocaine use	1	2.89E-07
		Current age	2	5.39E-10
		Recency of 1st opioid treatment	3	4.97E-06
		Cocaine use severity	4	4.69E-03
		N years using Heroin	5	2.80E-08
	SVM	Recency of last cocaine use	1	3.50E-08
		Recency of last cocaine injection	2	1.38E-04
		Current age	3	9.20E-07
		N years using Heroin	4	7.68E-06
		Recency of weekly cocaine use for 1 month	5	-
	RF	Recency of last cocaine use	1	4.38E-06
		Recency of weekly cocaine use for 1 month	2	-
		Used cocaine>11 last year	3	-
		Didn’t use cocaine>11 last year	4	-
		Recency of cocaine symptoms	5	-
	DNN	Recency of last cocaine use	1	1.04E-06
		Recency of last cocaine injection	2	-
		Recency of cocaine symptoms	3	1.11E-01
		Current age	4	5.50E-08
		Recency of weekly cocaine use for 1 month	5	-
Model 2	LASSO	Recency of last cocaine use	1	5.32E-15
		Current age	2	3.96E-08
		Cocaine use severity	3	1.24E-07
		Recency of last tobacco use	4	1.62E-03
		Recency first reported drug problems to professionals	5	9.65E-01
	SVM	Recency of last cocaine use	1	5.44E-07
		Recency of weekly cocaine use for 1 month	2	-
		Recency of cocaine symptoms	3	-

		Didn’t use cocaine>11 last year	4	2.61E-02
		Used cocaine>11 last year	5	5.16E-02
	RF	Recency of last cocaine use	1	2.26E-06
		Used cocaine>11 last year	2	-
		Didn’t use cocaine>11 last year	3	4.38E-02
		Recency of cocaine symptoms	4	-
		Recency of weekly cocaine use for 1 month	5	-
	DNN	Age first used tobacco	1	9.42E-04
		Had relationship for >1 year	2	6.54E-03
		Longest time in days without using tobacco	3	-
		Contacted relatives infrequently	4	-
		Mixed drugs with alcohol >3 times	5	-
Model 3	LASSO	Current age	1	5.20E-11
		HIV positive	2	8.55E-06
		Blamed others for one's mistake	3	1.76E-03
		Jobless for 6 month due to drugs/alcohol	4	2.93E-02
		Currently unemployed	5	2.97E-04
	SVM	Current age	1	2.67E-09
		N biological children	2	3.62E-02
		Currently unemployed	3	3.94E-02
		Age at maximum weight	4	-
		N months employed last year	5	-
	RF	Current age	1	2.49E-09
		Age at maximum weight	2	-
		BMI	3	5.98E-06
		N drug symptoms when having depression	4	1.13E-02
		current weight	5	-
	DNN	N doctor visits for health problems	1	-
		Current age	2	5.17E-11
		Jobless for 6 month due to drugs/alcohol	3	1.49E-01
		Treated for emotional, psychiatric or drug problems	4	3.51E-03
		Visited treatment center once last year	5	1.25E-01

European Americans

		Variable	Rank	p-value
Model 1	LASSO	Recency of last cocaine use	1	8.72E-19
		Current age	2	9.96E-12
		Recency started opioid treatment	3	9.31E-07
		Recency of last cocaine injection	4	1.80E-04
		Age last had antisocial behaviors	5	3.92E-06
	SVM	Recency of last cocaine use	1	1.01E-16
		Recency of last cocaine injection	2	1.01E-06
		Used cocaine>11 times in last year	3	3.10E-03
		Age at maximum weight	4	5.46E-02
		Current age	5	9.87E-03
	RF	Recency of last cocaine use	1	4.38E-06
		Recency of cocaine symptoms	2	3.24E-03
		Didn’t used cocaine>11 times in last year	3	2.79E-03
		Recency of last cocaine injection	4	6.52E-03
		Used cocaine>11 last year	5	-
	DNN	Recency of last cocaine use	1	4.79E-25
		Recency first brought up opioid problems with professionals	2	-
		Recency of last cocaine injection	3	1.07E-03
		Age last smoked cigarettes	4	6.60E-21
		No drunk driving arrests	5	-
Model 2	LASSO	Recency of last cocaine use	1	3.06E-18
		Current age	2	1.27E-05
		Recency of last cocaine injection	3	1.38E-08
		Age last had antisocial behaviors	4	4.56E-06
		Cocaine use severity	5	1.83E-07
	SVM	Recency of last cocaine use	1	3.14E-17

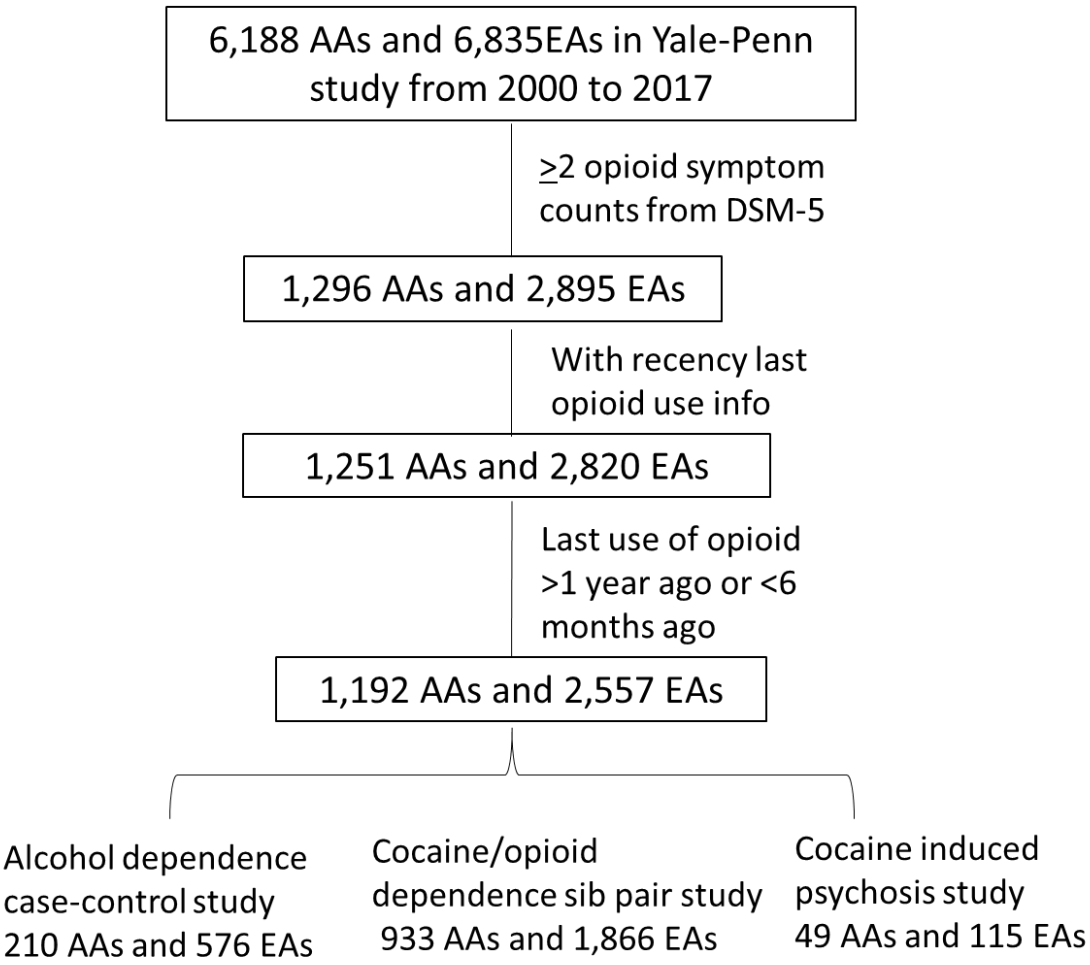
		Recency of last cocaine injection	2	1.74E-06
		Didn’t use cocaine>11 last year	3	8.36E-05
		Age at maximum weight	4	3.82E-02
		Current age	5	1.64E-05
	RF	Recency of last cocaine use	1	3.13E-14
		Recency of last cocaine injection	2	2.43E-06
		Used cocaine>11 last year	3	-
		Didn’t use cocaine>11 last year	4	8.16E-04
		Recency of cocaine symptoms	5	-
	DNN	Recency of last cocaine use	1	1.42E-17
		Didn’t hurt animal on purpose	2	-
		Recency last stayed high from cocaine >1 day	3	2.83E-03
		Recency had >2 cocaine symptoms	4	-
		Recency had cocaine symptoms	5	-
Model 3	LASSO	Current age	1	2.36E-09
		Current health has always been worse	2	7.62E-07
		BMI	3	3.12E-13
		Age last had antisocial behaviors	4	-
		Depression started with drug problems	5	3.98E-03
	SVM	Current age	1	3.10E-08
		Age at maximum weight	2	4.62E-03
		Age last had antisocial behaviors	3	2.74E-07
		BMI	4	5.89E-12
		Current weight	5	-
	RF	Current age	1	5.10E-09
		Age at maximum weight	2	2.36E-02
		BMI	3	1.15E-04
		Current weight	4	8.08E-02

		Max weight	5	4.51E-03
	DNN	Current age	1	2.13E-09
		Bad mood after ECT or bright light therapy	2	-
		Didn’t have OCD	3	-
		No gambling withdrawal when cannot gamble	4	9.20E-02
		Didn’t have sex with >10 people in a year	5	3.54 E-02

“Rank”: relative importance of a variable measured by the specified method; “-” : variable not selected by stepwise regression; LASSO: least absolute shrinkage and selection operator; SVM: support vector machine with recursive feature elimination; RF: random forest with recursive feature elimination; DNN: deep neural network with recursive feature elimination.

Appendix I Figure S1. Study subject derivation

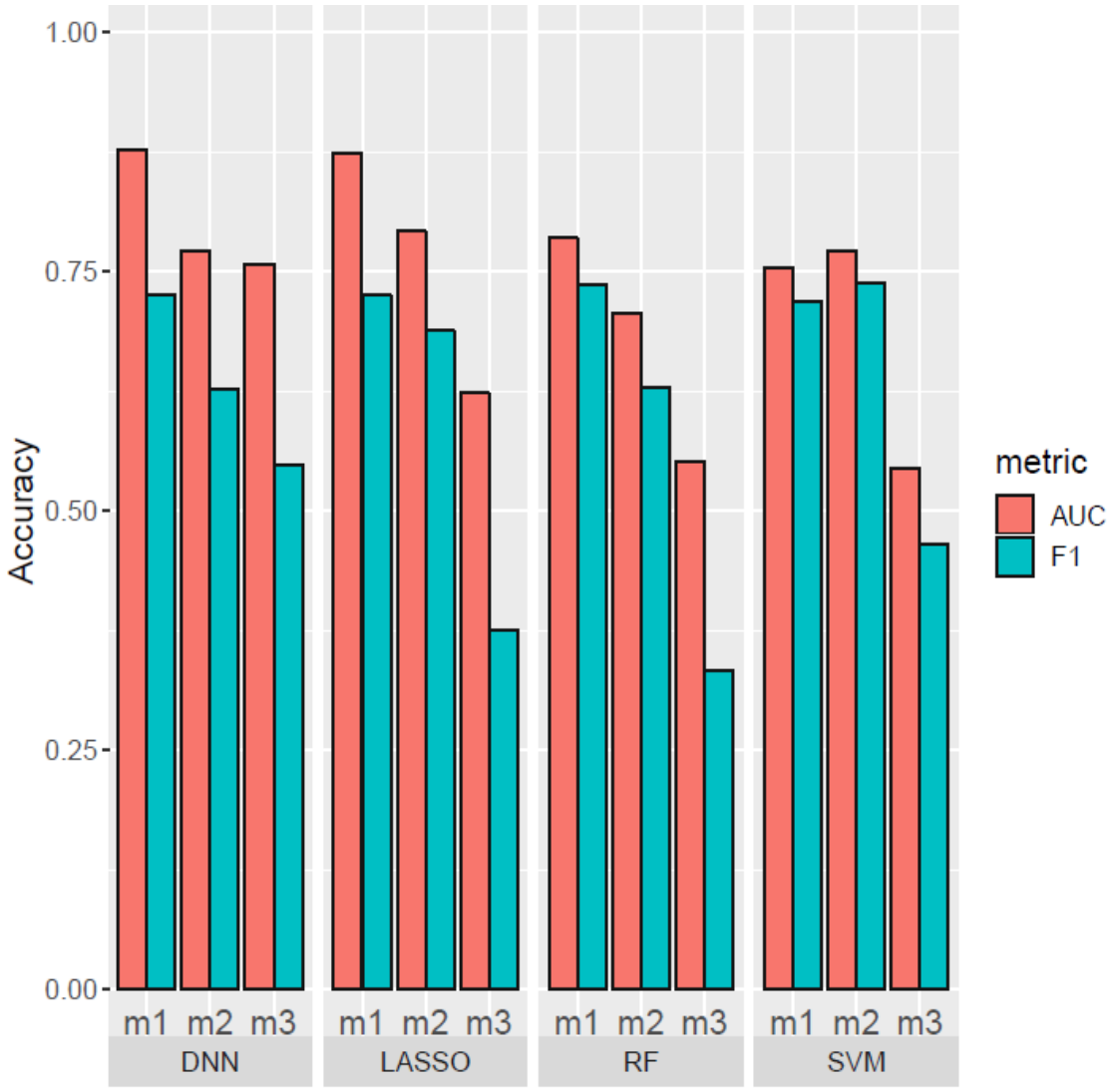
Derivation of African Americans (AAs) and European Americans (EAs) subjects in the Yale-Penn dataset who were ascertained from multiple substance use disorder studies and met criteria for cessation or non-cessation of opioid use.


6,188 AAs and 6,835EAs in Yale-Penn study from 2000 to 2017
≥2 opioid symptom counts from DSM-5
1,296 AAs and 2,895 EAs
With recency last opioid use info
1,251 AAs and 2,820 EAs
Last use of opioid >1 year ago or <6 months ago
1,192 AAs and 2,557 EAs
Alcohol dependence case-control study 210 AAs and 576 EAs
Cocaine/opioid dependence sib pair study 933 AAs and 1,866 EAs
Cocaine induced psychosis study 49 AAs and 115 EAs

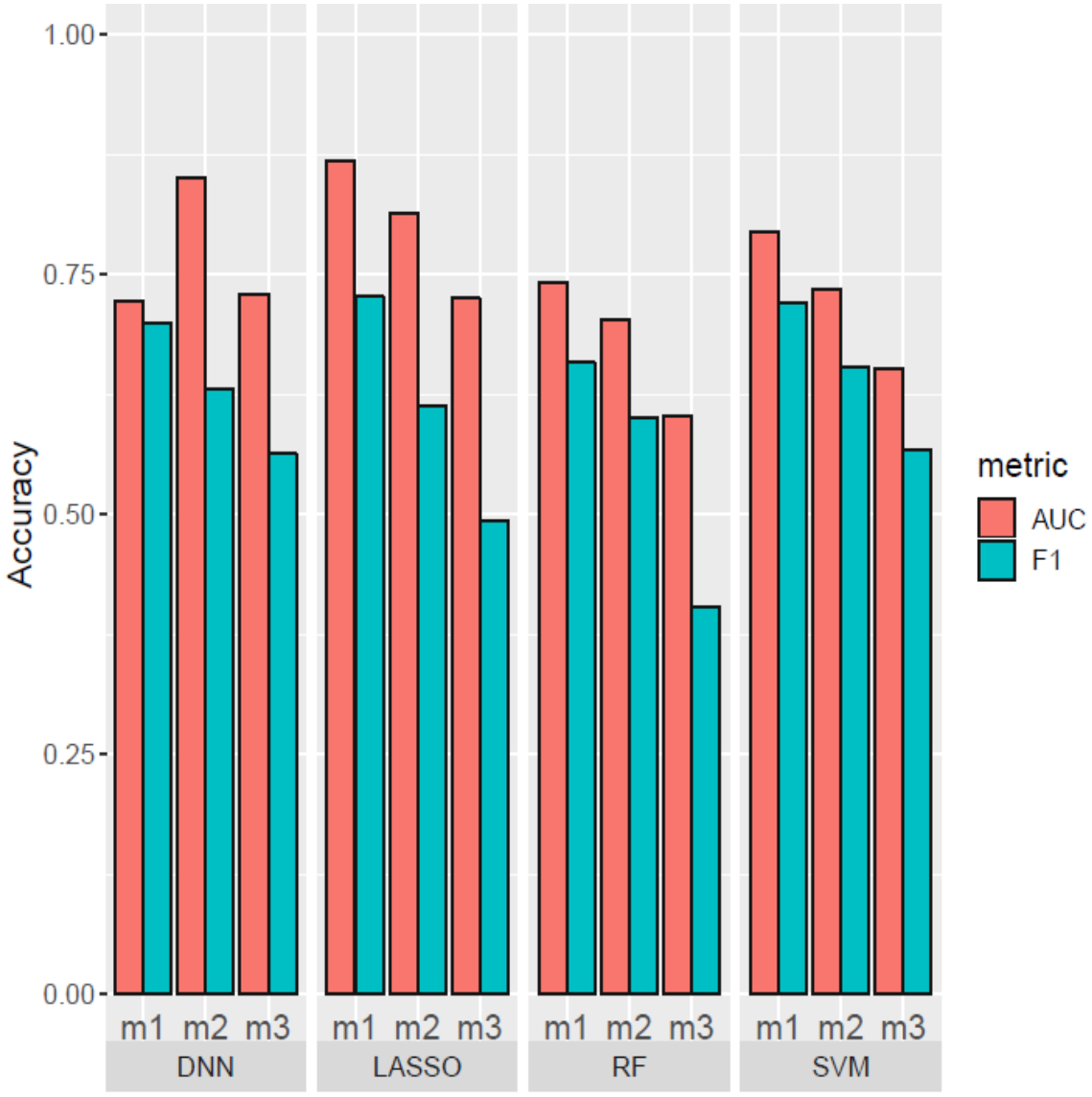
Appendix I Figure S2. Predictive accuracy of four machine learning methods by model in African Americans (A) and European Americans (B).

AUC and F1 scores derived from a common set of variables picked by the method (DNN, LASSO, RF, SVM) for each model. m1: model 1; m2: model 2; m3: model 3; DNN: deep neural network; LASSO: least absolute shrinkage and selection operator; RF: random forest; SVM: support vector machine; AUC: area under receiver operating curve; F1: F1 score.

A)


1.00
0.75
0.50
0.25
0.00
Accuracy
m1 m2 m3
DNN
LASSO
RF
SVM
metric
AUC
F1

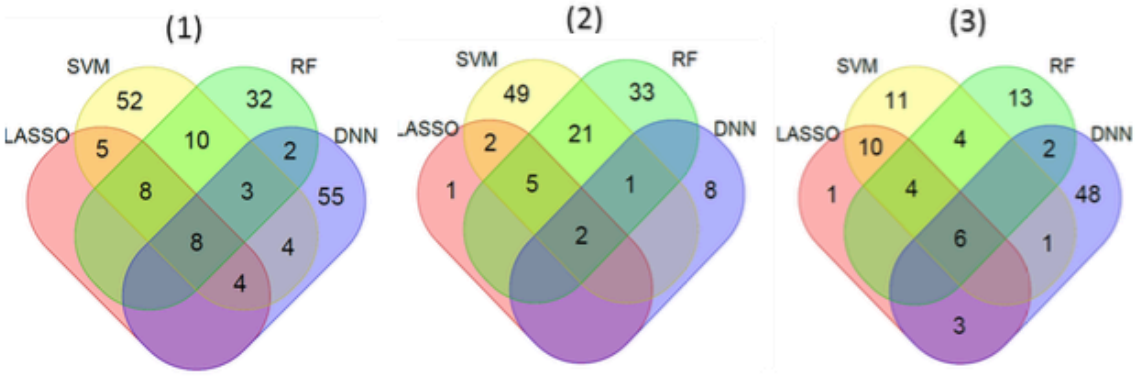
B)


1.00
0.75
0.50
0.25
0.00
Accuracy
m1 m2 m3
DNN
m1 m2 m3
LASSO
m1 m2 m3
RF
m1 m2 m3
SVM
metric
AUC
F1

Appendix I Figure S3. Number of overlapping "high impact" variables selected by each machine learning method based on the importance measurement in African Americans (A) and European Americans (B) for models (1), (2), and (3).

Model 1 includes all variables except for those that are confounded with opioid cessation. Model 2 includes all variables in Model 1 except opioid-related variables. Model 3 includes all variables in Model 2 except drug related variables. The criteria used for 'high impact' variables of each machine learning method can be found from **Materials and Methods** section in the main text. Colors represent machine learning method: pink = LASSO, light yellow = SVM, light green =RF, light blue = DNN.

A)


(1)
SVM
RF
LASSO
DNN
52
32
5
10
2
8
3
55
8
4
4
(2)
SVM
RF
LASSO
DNN
49
33
2
21
1
5
1
8
2
(3)
SVM
RF
LASSO
DNN
11
13
10
4
2
1
4
48
6
1
3

B)


(1)
SVM
RF
LASSO
DNN
28
41
1
1
2
5
7
3
7
1
4
1
1
(2)
SVM
RF
LASSO
DNN
40
47
1
6
3
2
1
60
1
7
2
(3)
SVM
RF
LASSO
DNN
23
17
3
13
1
2
9
24
1
5

APPENDIX II

Supplemental Methods for Chapter Four

Behavioral Testing

The behavioral tests were done previously by Lisa Goldberg as described[243], including OXY conditioned place preference in week 1 and 2, OXY induced locomotor activity in week 3, and elevated plus maze in week 4.

OXY conditioned place preference was done as described previously[283] using a 9-day protocol with 30-min test and training sessions. Initial side preference were assessed on Day 1 after animals received a saline injection and were allowed free access to both sides. On training days Day 2- Day 5, mice received either saline or OXY while being confined to the corresponding drug-paired side. On Day 8, conditioned place preference to the drug-paired side was assessed, as on Day 1. For locomotor activity measurement, a 3-day protocol was used[284-286]. On Day 1 and 2, mice were injected with SAL, and on Day 3 mice were injected with OXY, prior to immediate placement in the locomotor chamber. Activity was measured for 30-min post injection. Behavior in an elevated plus maze was assessed using a 5-day protocol. Mice received daily saline or OXY injections to induce dependence on Day 1- Day 4. On Day 5, emotional-affective component of spontaneous opioid withdrawal was measured in the elevated plus maze[96]. Percentage of time spent in the open arms [(open time)/(open time + closed time) *100)] was used as a primary metric: an increase in time spent in the open arms corresponds to opioid spontaneous withdrawal.

Tissue Collection

Striatal tissue was collected from 23 OXY-treated F_2 for mRNA sequencing 24 hours after elevated plus maze testing. Mice were chosen if they were either homozygous (J/J) or heterozygous (J/N) at the markers capturing a QTL on distal chromosome 1 for OXY-induced locomotor activity (rs6341208-163.13 Mb, rs51237371-181.32 Mb), which resulted in a sample size of 11-12 per genotype (3-4 females, 7-9 males per genotype). Brain tissue harvest, total RNA extraction and library preparations were done as previously described[286] by Lisa Goldberg. As a result, the 23 samples were run on 5 lanes (technical replicates) in parallel on an Illumina flow cell by Illumina HiSeq 4000 machine, an average of 69.4 million reads per sample was obtained. FASTQ files were quality checked via FASTQC, with Phred quality scores > 30 (i.e. less than 0.1% sequencing error).

Transcriptome Analysis

Using Rsubread[287], FASTQ files were used to align reads to the reference genome (mm10; UCSC Genome Browser) and read counts per gene were quantified. Genes with minimum of one read across all quadruplicates as well as a minimum of one count per million in at least 25% of the count files were included. The read counts of each gene were normalized by VOOM transformation[245] using R package limma[245]. For each sample, we regressed the gene expression level to the Genotype (across the Chr 1 QTL interval), adjusting

for sex; sample replicate was treated as a random effects measure to estimate within-sample correlation.

Trans-eQTL Analysis

Differential gene expression in response to trans-located markers were done as previously described[243]. FASTQ files from each sample were summed across technical replicates and aligned to reference genome (mm10; UCSC Genome Browser) using Rsubread[288]. Genes with at least one fragment per kilobase of transcript per million mapped reads across all samples were used for the analysis. Expression levels of 12739 genes were regressed to 90 SNPs using a linear model with an additive model in R package Matrix eQTL[246]. Trans-eQTL was defined as any SNP in the panel that regulated genes on a different chromosome through expression. Genes possessing significant trans-eQTL were defined by a FDR cut off of 0.5%, which yielded 3495 genes.

BIBLIOGRAPHY

1. Phillips JK, Ford, MA, Bonnie RJ (eds.) *Pain Management and the Opioid Epidemic: Balancing Societal and Individual Benefits and Risks of Prescription Opioid Use.* National Academies Press (US) (2017).

2. Kolodny, A. *et al.* The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. *Annual Review of Public Health* **36**, 559-74 (2015).

3. CDC. (2017).

4. SAMHSA. Results from the 2012 National Survey on Drug Use and Health: Summary of national findings. Rockville, MD. *2013 National Survey of Substance Abuse Treatment* **(NSDUH Series H-46)**.

5. National Centers for Disease Control and Prevention, N.C.f.I. & Prevention and Control, D.o.U.I.P. (2016).

6. Compton, W.M., Jones, C.M. & Baldwin, G.T. Relationship between Nonmedical Prescription-Opioid Use and Heroin Use. *The New England Journal of Medicine* **374**, 154-163 (2016).

7. Moberg, K. The Role of Managed Care Professionals and Pharmacists in Combating Opioid Abuse. *Combating the Opioid Epidemic* (2018).

8. Robin Ghertner , L.G. The Opioid Crisis and Economic Opportunity: Geographic and Economic Trends. *ASPE reserach Brief, U.S. Department of Health and Human Services* (2018).

9. Ho, J.Y. The Contribution of Drug Overdose to Educational Gradients in Life Expectancy in the United States, 1992-2011. *Demography* **54**, 1175-1202 (2017).

10. Krueger, A.B. Where Have All the Workers Gone? An Inquiry into the Decline of the U.S. Labor Force Participation Rate. *Brookings Papers on Economic Activity* **2017**, 1-87 (2017).

11. Pathan, H. & Williams, J. Basic opioid pharmacology: an update. *British Journal of Pain* **6**, 11-16 (2012).

12. Dhawan, B.N. *et al.* International Union of Pharmacology. XII. Classification of opioid receptors. *Pharmacological Reviews* **48**, 567-592 (1996).

13. Chahl, L.A. Opioids - mechanisms of action. *Australian Prescriber* **19,** 63–65 (1996).

14. Al-Hasani, R. & Bruchas, M.R. Molecular mechanisms of opioid receptor-dependent signaling and behavior. *Anesthesiology* **115**, 1363-1381 (2011).

15. Raman, M., Chen, W. & Cobb, M.H. Differential regulation and properties of MAPKs. *Oncogene* **26**, 3100-3112 (2007).

16. Le Merrer, J., Becker, J.A., Befort, K. & Kieffer, B.L. Reward processing by the opioid system in the brain. *Physiological Reviews* **89**, 1379-1412 (2009).

17. Vaughan, C.W., Ingram, S.L., Connor, M.A. & Christie, M.J. How opioids inhibit GABA-mediated neurotransmission. *Nature* **390**, 611-614 (1997).

18. Mahesh Trivedi*, S.S., Carl Gwinnut. Pharmacology of Opioids. *Anaesthesia Tutorial of the Week* (2007).

19. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). *National Center for Health Statistics* (2015).

20. Arlington, V. Diagnostic and statistical manual of mental disorders (5th ed.). *American Psychiatric Association.* (2013).

21. Hasin, D.S. *et al.* DSM-5 criteria for substance use disorders: recommendations and rationale. *American Journal of Psychiatry* **170**, 834-851 (2013).

22. Organization, W.H. Annex 3ICD-10 codes for conditions covered in these guidelines. *Guidelines for the Psychosocially Assisted Pharmacological Treatment of Opioid Dependence.* (2009).

23. Organization, W.H. international statistical classification of diseases and related health problems : tenth revision. 2nd ed. (2004).

24. Coffin, P.O. *et al.* Behavioral intervention to reduce opioid overdose among high-risk persons with opioid use disorder: A pilot randomized controlled trial. *PLoS One* **12**, e0183354 (2017).

25. McAuliffe, W.E. & Ch'ien, J.M. Recovery training and self help: a relapse-prevention program for treated opiate addicts. *Journal of Substance Abuse Treatment* **3**, 9-20 (1986).

26. Ali, S., Tahir, B., Jabeen, S. & Malik, M. Methadone Treatment of Opiate Addiction: A Systematic Review of Comparative Studies. *Innovations in Clinical Neuroscience* **14**, 8-19 (2017).

27. Garcia-Portilla, M.P., Bobes-Bascaran, M.T., Bascaran, M.T., Saiz, P.A. & Bobes, J. Long term outcomes of pharmacological treatments for opioid dependence: does methadone still lead the pack? *British Journal of Clinical Pharmacology* **77**, 272-284 (2014).

28. Schwartz, R.P. *et al.* A randomized controlled trial of interim methadone maintenance. *Archives of General Psychiatry* **63**, 102-109 (2006).

29. Kosten, T.R. & George, T.P. The neurobiology of opioid dependence: implications for treatment. *Addiction Science & Clinical Practice* **1**, 13-20 (2002).

30. Gelernter, J. & Kranzler, H.R. Genetics of drug dependence. *Dialogues in Clinical Neuroscience* **12**, 77-84 (2010).

31. Gelernter, J. *et al.* Genomewide linkage scan for opioid dependence and related traits. *American Journal of Medical Genetics* **78**, 759-769 (2006).

32. Glatt, S.J. *et al.* Genome-wide linkage analysis of heroin dependence in Han Chinese: results from Wave Two of a multi-stage study. *Drug and Alcohol Dependence* **98**, 30-34 (2008).

33. Patnala, R., Clements, J. & Batra, J. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genetics* **14**, 39 (2013).

34. Beer, B. *et al.* Association of polymorphisms in pharmacogenetic candidate genes (OPRD1, GAL, ABCB1, OPRM1) with opioid dependence in European population: a case-control study. *PLoS One* **8**, e75359 (2013).

35. Gao, X. *et al.* Contribution of Genetic Polymorphisms and Haplotypes in DRD2, BDNF, and Opioid Receptors to Heroin Dependence and Endophenotypes Among the Han Chinese. *OMICS* **21**, 404-412 (2017).

36. Bawor, M. *et al.* Contribution of BDNF and DRD2 genetic polymorphisms to continued opioid use in patients receiving methadone treatment for opioid use disorder: an observational study. *Addiction Science & Clinical Practice* **10**, 19 (2015).

37. Hayes, B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). *Methods in Molecular Biology* **1019**, 149-169 (2013).

38. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* **101**, 5-22 (2017).

39. Gelernter, J. *et al.* Genome-wide association study of opioid dependence: multiple associations mapped to calcium and potassium pathways. *Biological Psychiatry* **76**, 66-74 (2014).

40. Cheng, Z. *et al.* Genome-wide Association Study Identifies a Regulatory Variant of RGMA Associated With Opioid Dependence in European Americans. *Biological Psychiatry* **84**, 762-770 (2018).

41. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews. Genetics* **13**, 135-145 (2012).

42. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).

43. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics* **95**, 5-23 (2014).

44. Lee, H.K., Braynen, W., Keshav, K. & Pavlidis, P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* **6**, 269 (2005).

45. Irizarry, R.A., Wang, C., Zhou, Y. & Speed, T.P. Gene set enrichment analysis made simple. *Statistical Methods in Medical Research* **18**, 565-575 (2009).

46. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90-97 (2016).

47. Delker, E., Brown, Q. & Hasin, D. Epidemiological Studies of Substance Dependence and Abuse in Adults. *Current Behavioral Neuroscience Reports* **2**, 15-22 (2015).

48. Grant, B.F. *et al.* Epidemiology of DSM-5 Drug Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions-III. *JAMA Psychiatry* **73**, 39-47 (2016).

49. E Cabrina Campbell, M.N. Caroff, Stephan C Mann. Co-occurring schizophrenia and substance use disorder: Epidemiology, pathogenesis, clinical manifestations, course, assessment and diagnosis. UpToDate® (2017).

50. Birhanu, A.M., Bisetegn, T.A. & Woldeyohannes, S.M. High prevalence of substance use and associated factors among high school adolescents in Woreta Town, Northwest Ethiopia: multi-domain factor analysis. *BMC Public Health* **14**, 1186 (2014).

51. Horyniak, D., Melo, J.S., Farrell, R.M., Ojeda, V.D. & Strathdee, S.A. Epidemiology of Substance Use among Forced Migrants: A Global Systematic Review. *PLoS One* **11**, e0159134 (2016).

52. Sirratt, D., Ozanian, A. & Traenkner, B. Epidemiology and prevention of substance use disorders in the military. *Military Medicine* **177**, 21-28 (2012).

53. Zhang, Z. Too much covariates in a multivariable model may cause the problem of overfitting. *Journal of Thoracic Disease* **6**, E196-197 (2014).

54. A. Jović, K. Brkic, N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* 1200–1205 (2015). https://doi.org/10.1109/MIPRO.2015.7160458

55. Sanz, H., Valim, C., Vegas, E., Oller, J.M. & Reverter, F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* **19**, 432 (2018).

56. Smith, G. Step away from stepwise. *Journal of Big Data* **5**(2018).

57. Ranganathan, P., Pramesh, C.S. & Aggarwal, R. Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research* **8**, 148-151 (2017).

58. Connor, P., Hollensen, P., Krigolson, O. & Trappenberg, T. A biological mechanism for Bayesian feature selection: Weight decay and raising the LASSO. *Neural Networks* **67**, 121-130 (2015).

59. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).

60. Tibshirani, R.J. The lasso problem and uniqueness. *Electronic Journal of Statistics* **7**, 1456-1490 (2013).

61. Liu, J., Ranka, S. & Kahveci, T. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics* **24**, i86-95 (2008).

62. Bron, E.E., Smits, M., Niessen, W.J. & Klein, S. Feature Selection Based on the SVM Weight Vector for Classification of Dementia. *IEEE Journal of Biomedical and Health Informatics* **19**, 1617-1626 (2015).

63. Isabelle Guyon , J.W., Stephen Barnhill. Gene selection for cancer classification using support vector machines. *8th International Conference on Machine Learning and Data Mining in Pattern Recognition,* **46**, 389-422 (2002).

64. Menze, B.H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009).

65. Rania Saidi, W.B.N., Nadia Essoussi. Feature Selection Using Genetic Algorithm for Big Data. *International Conference on Advanced Machine Learning Technologies and Applications*, 352-361 (2018).

66. Sabah Sayed, M.N., fAmr Badr, Ibrahim Farag. A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets. *Expert Systems with Applications* **121**, 233-243 (2019).

67. Schafer, W.R. Addiction research in a simple animal model: the nematode Caenorhabditis elegans. *Neuropharmacology* **47 Suppl 1**, 123-31 (2004).

68. Kaun, K.R., Devineni, A.V. & Heberlein, U. Drosophila melanogaster as a model to study drug addiction. *Human Genetics* **131**, 959-75 (2012).

69. Klee, E.W. *et al.* Zebrafish: a model for the study of addiction genetics. *Human Genetics* **131**, 977-1008 (2012).

70. Mott, R. & Flint, J. Dissecting quantitative traits in mice. *Annual Review of Genomics and Human Genetics* **14**, 421-39 (2013).

71. Flint, J. & Eskin, E. Genome-wide association studies in mice. *Nature reviews. Genetics* **13**, 807-17 (2012).

72. Lin, X. & Wells, D. Isolation of the mouse cDNA homologous to the human EXT1 gene responsible for Hereditary Multiple Exostoses. *DNA Sequence* **7**, 199-202 (1997).

73. Kile, B.T. & Hilton, D.J. The art and design of genetic screens: mouse. *Nature Reviews. Genetics,* **6**, 557-67 (2005).

74. Cryan, J.F. & Holmes, A. The ascent of mouse: advances in modelling human depression and anxiety. *Nature Reviews. Drug Discovery* **4**, 775-90 (2005).

75. Edwards, S. & Koob, G.F. Experimental psychiatric illness and drug abuse models: from human to animal, an overview. *Methods in Molecular Biology* **829**, 31-48 (2012).

76. Koob, G.F. & Volkow, N.D. Neurocircuitry of addiction. *Neuropsychopharmacology* **35**, 217-38 (2010).

77. Nemudryi, A.A., Valetdinova, K.R., Medvedev, S.P. & Zakian, S.M. TALEN and CRISPR/Cas Genome Editing Systems: Tools of Discovery. *Acta Naturae* **6**, 19-40 (2014).

78. Hartenian, E. & Doench, J.G. Genetic screens and functional genomics using CRISPR/Cas9 technology. *FEBS Journal* **282**, 1383-93 (2015).

79. Gatti, D.M. *et al.* Quantitative trait locus mapping methods for diversity outbred mice. *Genes, Genomes, Genetics (Bethesda)* **4**, 1623-33 (2014).

80. Berrettini, W.H., Ferraro, T.N., Alexander, R.C., Buchberg, A.M. & Vogel, W.H. Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. *Nature Genetics* **7**, 54-8 (1994).

81. Metten, P., Crabbe, J.C. & Belknap, J.K. Genetic correlates of morphine withdrawal in 14 inbred mouse strains. *Drug and Alcohol Dependence* **99**, 123-31 (2009).

82. Bryant, C.D. *et al.* A major QTL on chromosome 11 influences psychostimulant and opioid sensitivity in mice. *Genes, Brain and Behavior* **8**, 795-805 (2009).

83. Crusio, W.E., Dhawan, E., Chesler, E.J. & Delprato, A. Analysis of morphine responses in mice reveals a QTL on Chromosome 7. *F1000Rearch* **5**, 2156 (2016).

84. Loo, L. *et al.* Single-cell transcriptomic analysis of mouse neocortical development. *Nature Communications* **10**, 134 (2019).

85. Poulin, J.F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J.M. & Awatramani, R. Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience* **19**, 1131-41 (2016).

86. Moroy, T. & Heyd, F. The impact of alternative splicing in vivo: mouse models show the way. *RNA* **13**, 1155-71 (2007).

87. Avey, D. *et al.* Single-Cell RNA-Seq Uncovers a Robust Transcriptional Response to Morphine by Glia. *Cell Reports* **24**, 3619-3629 e4 (2018).

88. Zhou, Z., Enoch, M.A. & Goldman, D. Gene expression in the addicted brain. *International Review of Neurobiology* **116**, 251-73 (2014).

89. Grimes, T., Potter, S.S. & Datta, S. Integrating gene regulatory pathways into differential network analysis of gene expression data. *Science Reports* **9**, 5479 (2019).

90. Reyes-Gibby, C.C., Yuan, C., Wang, J., Yeung, S.C. & Shete, S. Gene network analysis shows immune-signaling and ERK1/2 as novel genetic markers for multiple addiction phenotypes: alcohol, smoking and opioid addiction. *BMC Systems Biology* **9**, 25 (2015).

91. Corey, P.J., Heck, A.M. & Weathermon, R.A. Amphetamines to counteract opioid-induced sedation. *Annals of Pharmacotherapy* **33**, 1362-6 (1999).

92. Zhang, J.J. & Kong, Q. Locomotor activity: A distinctive index in morphine self-administration in rats. *PLoS One* **12**, e0174272 (2017).

93. Patti, C.L. *et al.* Behavioral characterization of morphine effects on motor activity in mice. *Pharmacology Biochemistry and Behavior* **81**, 923-7 (2005).

94. Runegaard, A.H. *et al.* Locomotor- and Reward-Enhancing Effects of Cocaine Are Differentially Regulated by Chemogenetic Stimulation of Gi-Signaling in Dopaminergic Neurons. *Neurology* **5**(2018).

95. LeroyHoffman, K. 3 - Modeling disorders of fear and anxiety in animals. *Modeling Neuropsychiatric Disorders in Laboratory Animals*, 87-160 (2016).

96. Schulteis, G., Yackey, M., Risbrough, V. & Koob, G.F. Anxiogenic-like effects of spontaneous and naloxone-precipitated opiate withdrawal in the elevated plus-maze. *Pharmacology, Biochemistry, and Behavior* **60**, 727-31 (1998).

97. Buckman, S.G., Hodgson, S.R., Hofford, R.S. & Eitan, S. Increased elevated plus maze open-arm time in mice during spontaneous morphine withdrawal. *Behavioural Brain Research* **197**, 454-456 (2009).

98. Hodgson, S.R., Hofford, R.S., Norris, C.J. & Eitan, S. Increased elevated plus maze open-arm time in mice during naloxone-precipitated morphine withdrawal. *Behavioural Pharmacology* **19**, 805-11 (2008).

99. Dasgupta, N., Beletsky, L. & Ciccarone, D. Opioid Crisis: No Easy Fix to Its Social and Economic Determinants. *American Journal of Public Health* **108**, 182-186 (2018).

100. CDC. Centers for Disease Control and Prevention. Overview of the drug overdose epidemic: behind the numbers. (2019). https://www.cdc.gov/drugoverdose/data/index.html

101. U.S. Department of Health and Human Services. What is the U.S. opioid epidemic? (2019). https://www.hhs.gov/opioids/about-the-epidemic/index.html

102. Samet, J.H., Botticelli, M. & Bharel, M. Methadone in Primary Care - One Small Step for Congress, One Giant Leap for Addiction Treatment. *New England Journal of Medicine* **379**, 7-8 (2018).

103. Schuckit, M.A. Treatment of Opioid-Use Disorders. *The New England Journal of Medicine* **375**, 357-368 (2016).

104. O'Connor, P.G. Methods of detoxification and their role in treating patients with opioid dependence. *JAMA: The Journal of the American Medical Association* **294**, 961-963 (2005).

105. Cheng, Z. *et al.* Genome-wide Association Study Identifies a Regulatory Variant of RGMA Associated With Opioid Dependence in European Americans. *Biological Psychiatry* **84**, 762-770 (2018).

106. Mistry, C.J., Bawor, M., Desai, D., Marsh, D.C. & Samaan, Z. Genetics of Opioid Dependence: A Review of the Genetic Contribution to Opioid Dependence. *Current Psychiatry Reviews* **10**, 156-167 (2014).

107. Scott, C.K., Dennis, M.L., Laudet, A., Funk, R.R. & Simeone, R.S. Surviving drug addiction: the effect of treatment and abstinence on mortality. *American Journal of Public Health* **101**, 737-744 (2011).

108. Tkacz, J., Severt, J., Cacciola, J. & Ruetsch, C. Compliance with buprenorphine medication-assisted treatment and relapse to opioid use. *The American Journal on Addictions* **21**, 55-62 (2012).

109. Yang, B.Z., Han, S., Kranzler, H.R., Palmer, A.A. & Gelernter, J. Sex-specific linkage scans in opioid dependence. *American Journal of Medical Genetics* **174**, 261-268 (2017).

110. Nelson, E.C. *et al.* Evidence of CNIH3 involvement in opioid dependence. *Molecular Psychiatry* **21**, 608-14 (2016).

111. Rapid GWAS of thousands of phenotypes for 337,000 samples in the UK biobank (2017). http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank

112. Gelernter, J. *et al.* Genomewide linkage scan for cocaine dependence and related traits: significant linkages for a cocaine-related trait and cocaine-induced paranoia. *American Journal of Medical Genetics* **136B**, 45-52 (2005).

113. Sherva, R. *et al.* Genome-wide Association Study of Cannabis Dependence Severity, Novel Risk Variants, and Shared Genetic Risks. *JAMA Psychiatry* **73**, 472-80 (2016).

114. Pierucci-Lagha, A. *et al.* Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug and Alcohol Dependence* **91**, 85-90 (2007).

115. Malison, R.T. *et al.* Inter-rater reliability and concurrent validity of DSM-IV opioid dependence in a Hmong isolate using the Thai version of the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Addictive Behaviors* **36**, 156-60 (2011).

116. Nelson, E.C. *et al.* Association of OPRD1 polymorphisms with heroin dependence in a large case-control series. *Addiction Biology* **19**, 111-21 (2014).

117. Bucholz, K.K. *et al.* A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *Journal of Studies on Alcohol and Drugs* **55**, 149-58 (1994).

118. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **81**, 559-75 (2007).

119. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-9 (2006).

120. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955-9 (2012).

121. Shaun Purcell, C.C. Plink 1.9.

122. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

123. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

124. Purcell, S., Cherny, S.S. & Sham, P.C. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19**, 149-50 (2003).

125. Carithers, L.J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking* **13**, 311-9 (2015).

126. Ramasamy, A. *et al.* Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature Neuroscience* **17**, 1418-1428 (2014).

127. Kundu, S., Aulchenko, Y.S., van Duijn, C.M. & Janssens, A.C. PredictABEL: an R package for the assessment of risk prediction models. *European Journal of Epidemiology* **26**, 261-4 (2011).

128. Waljee, J.F. & Brummett, C.M. Opioid Prescribing for Low Back Pain: What Is the Role of Payers? *JAMA Network Open* **1**, e180236 (2018).

129. Deyo, R.A., Von Korff, M. & Duhrkoop, D. Opioids for low back pain. *BMJ: British Medical Journal* **350**, g6380 (2015).

130. Euesden, J., Lewis, C.M. & O'Reilly, P.F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466-8 (2015).

131. Segre, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genetics* **6**(2010).

132. Uhl, G.R. *et al.* Cocaine reward is reduced by decreased expression of receptor-type protein tyrosine phosphatase D (PTPRD) and by a novel PTPRD antagonist. *Proceedings of the National Academy of Sciences of the United States of America* **115**, 11597-11602 (2018).

133. Meehan, M. *et al.* Protein tyrosine phosphatase receptor delta acts as a neuroblastoma tumor suppressor by destabilizing the aurora kinase A oncogene. *Molecular Cancer* **11**, 6 (2012).

134. Lind, P.A. *et al.* A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations. *Twin Research and Human Genetics* **13**, 10-29 (2010).

135. Corradini, I., Verderio, C., Sala, M., Wilson, M.C. & Matteoli, M. SNAP-25 in neuropsychiatric disorders. *Annals of the New York Academy of Sciences* **1152**, 93-99 (2009).

136. Muller, D.J. *et al.* The SNAP-25 gene may be associated with clinical response and weight gain in antipsychotic treatment of schizophrenia. *Neuroscience Letters* **379**, 81-89 (2005).

137. Wang, X. *et al.* Phosphorylated SNAP25 in the CA1 regulates morphine-associated contextual memory retrieval via increasing GluN2B-NMDAR surface localization. *Addiction Biology* **23**, 1067-1078 (2018).

138. Feng, Y. *et al.* The SNAP25 gene as a susceptibility gene contributing to attention-deficit hyperactivity disorder. *Molecular Psychiatry* **10**, 998-1005, 973 (2005).

139. KalleianEserian, J. Vitamin D as an effective treatment approach for drug abuse and addiction. *Journal of Medical Hypotheses and Ideas* **7**, 35-39 (2013).

140. Poisbeau P, A.M., Gazzo G. Cholecalciferol (Vitamin D3) Reduces Rat Neuropathic Pain by Modulating Opioid Signaling. *Molecular Neurobiology* (2019).

141. Kim, T.W., Alford, D.P., Holick, M.F., Malabanan, A.O. & Samet, J.H. Low vitamin d status of patients in methadone maintenance treatment. *Journal of Addiction Medicine* **3**, 134-138 (2009).

142. Bergman, P., Sperneder, S., Hoijer, J., Bergqvist, J. & Bjorkhem-Bergman, L. Low vitamin D levels are associated with higher opioid dose in palliative cancer patients--results from an observational study in Sweden. *PLoS One* **10**, e0128223 (2015).

143. Turner, M.K. *et al.* Prevalence and clinical correlates of vitamin D inadequacy among patients with chronic pain. *Pain Medicine* **9**, 979-984 (2008).

144. Gudin, J. Opioid therapies and cytochrome p450 interactions. *Journal of Pain Symptom Management* **44**, S4-14 (2012).

145. Leppert, W. CYP2D6 in the metabolism of opioids for mild to moderate pain. *Pharmacology* **87**, 274-285 (2011).

146. Yiannakopoulou, E. Pharmacogenomics and Opioid Analgesics: Clinical Implications. *International Journal of Genomics* **2015**, 368979 (2015).

147. Ahmed, M., Ul Haq, I., Faisal, M., Waseem, D. & Taqi, M.M. Implication of OPRM1 A118G Polymorphism in Opioids Addicts in Pakistan: In vitro and In silico Analysis. *Journal of Molecular Neuroscience* **65**, 472-479 (2018).

148. Wrzosek, M. *et al.* Association between Fok I vitamin D receptor gene (VDR) polymorphism and impulsivity in alcohol-dependent patients. *Molecular Biology Reports* **41**, 7223-7228 (2014).

149. Fujita-Hamabe, W., Nakamoto, K. & Tokuyama, S. Involvement of NCAM and FGF receptor signaling in the development of analgesic tolerance to morphine. *European Journal of Pharmacology* **672**, 77-82 (2011).

150. Belcheva, M.M., Haas, P.D., Tan, Y., Heaton, V.M. & Coscia, C.J. The fibroblast growth factor receptor is at the site of convergence between mu-opioid receptor and growth factor signaling pathways in rat C6 glioma cells. *Journal of Pharmacology and Experimental Therapeutics* **303**, 909-918 (2002).

151. Skupio, U. *et al.* Behavioral and transcriptional patterns of protracted opioid self-administration in mice. *Addiction Biology* **22**, 1802-1816 (2017).

152. Beenken, A. & Mohammadi, M. The FGF family: biology, pathophysiology and therapy. *Nature Reviews. Drug Discovery* **8**, 235-253 (2009).

153. Isakova, T. *et al.* Associations of FGF23 With Change in Bone Mineral Density and Fracture Risk in Older Individuals. *Journal of Bone and Mineral Research* **31**, 742-748 (2016).

154. Coluzzi, F., Pergolizzi, J., Raffa, R.B. & Mattia, C. The unsolved case of "bone-impairing analgesics": the endocrine effects of opioids on bone metabolism. *Therapeutics and Clinical Risk Management* **11**, 515-523 (2015).

155. Tsuang, M.T. *et al.* Co-occurrence of abuse of different drugs in men: the role of drug-specific and shared vulnerabilities. *Archives of General Psychiatry* **55**, 967-972 (1998).

156. Vink, J.M. *et al.* Polygenic risk scores for smoking: predictors for alcohol and cannabis use? *Addiction* **109**, 1141-1151 (2014).

157. Dennis, B.B. *et al.* Pain and Opioid Addiction: A Systematic Review and Evaluation of Pain Measurement in Patients with Opioid Dependence on Methadone Maintenance Treatment. *Current Drug Abuse Reviews* **9**, 49-60 (2016).

158. Busse, J.W. *et al.* Opioids for Chronic Noncancer Pain: A Systematic Review and Meta-analysis. *JAMA: The Journal of the American Medical Association* **320**, 2448-2460 (2018).

159. Liang, D.Y. *et al.* The Netrin-1 receptor DCC is a regulator of maladaptive responses to chronic morphine administration. *BMC Genomics* **15**, 345 (2014).

160. Suri, P. *et al.* Genome-wide meta-analysis of 158,000 individuals of European ancestry identifies three loci associated with chronic back pain. *PLoS Genetics* **14**, e1007601 (2018).

161. Andersen, A.M. *et al.* Polygenic Scores for Major Depressive Disorder and Risk of Alcohol Dependence. *JAMA Psychiatry* **74**, 1153-1160 (2017).

162. Volkow, N.D., Frieden, T.R., Hyde, P.S. & Cha, S.S. Medication-assisted therapies--tackling the opioid-overdose epidemic. *New England Journal of Medicine* **370**, 2063-6 (2014).

163. Acion, L. *et al.* Use of a machine learning framework to predict substance use disorder treatment success. *PLoS One* **12**, e0175383 (2017).

164. Che, Z., St Sauver, J., Liu, H. & Liu, Y. Deep Learning Solutions for Classifying Patients on Opioid Use. *AMIA 2019 Annual Symposium* **2017**, 525-534 (2017).

165. Diagnostic and Statistical Manual of Metal Disorders, Fifth Edition.

166. James, K. & Jordan, A. The Opioid Crisis in Black Communities. *The Journal of Law, Medicine & Ethics* **46**, 404-421 (2018).

167. Hansen, H., Siegel, C., Wanderling, J. & DiRocco, D. Buprenorphine and methadone treatment for opioid dependence by income, ethnicity and race

of neighborhoods in New York City. *Drug and Alcohol Dependence* **164**, 14-21 (2016).

168. Om, A. The opioid crisis in black and white: the role of race in our nation's recent drug epidemic. *Journal of Public Health* **40**, e614-e615 (2018).

169. Santoro, T.N. & Santoro, J.D. Racial Bias in the US Opioid Epidemic: A Review of the History of Systemic Bias and Implications for Care. *Cureus* **10**, e3733 (2018).

170. Hellman, M. Opioids, opioids, opioids: The plague among middle-aged white Americans. *Nordic Studies on Alcohol and Drugs* **35**(2018).

171. Badiani, A. & Spagnolo, P.A. Role of environmental factors in cocaine addiction. *Current Pharmaceutical Design* **19**, 6996-7008 (2013).

172. Randall J. Ellis, Z.W., Nicholas Genes and Avi Ma'ayan. Predicting opioid dependence from electronic health records with machine learning. *BioData Mining* **12**(2019).

173. LiX, C., CurranG, TilfordJM, FelixH1MartinBC. PSY5 - Using Machine Learning to Predict Opioid Overdoses Among Prescription Opioid Users. *Value in Health* **21**, S245 (2018).

174. Dowell, D., Haegerich, T.M. & Chou, R. CDC Guideline for Prescribing Opioids for Chronic Pain - United States, 2016. *MMWR Recommendations and Reports* **65**, 1-49 (2016).

175. Ainscough, T.S., Brose, L.S., Strang, J. & McNeill, A. Contingency management for tobacco smoking during opioid addiction treatment: a randomised pilot study. *BMJ Open* **7**, e017467 (2017).

176. Samet, J.H. & Kertesz, S.G. Suggested Paths to Fixing the Opioid Crisis: Directions and Misdirections. *JAMA Network Open* **1**, e180218 (2018).

177. Pierucci-Lagha, A. *et al.* Reliability of DSM-IV diagnostic criteria using the semi-structured assessment for drug dependence and alcoholism (SSADDA). *Drug and Alcohol Dependence* **91**, 85-90 (2007).

178. Montalvo-Ortiz, J.L., Cheng, Z., Kranzler, H.R., Zhang, H. & Gelernter, J. Genomewide Study of Epigenetic Biomarkers of Opioid Dependence in European- American Women. *Science Reports* **9**, 4660 (2019).

179. Wetherill, L. *et al.* Association of substance dependence phenotypes in the COGA sample. *Addiction Biology* **20**, 617-627 (2015).

180. Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nature Methods* **15**, 233-234 (2018).

181. Beam, A.L. & Kohane, I.S. Big Data and Machine Learning in Health Care. *JAMA: The Journal of the American Medical Association* **319**, 1317-1318 (2018).

182. Wong, A. *et al.* Development and Validation of an Electronic Health Record-Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. *JAMA Network Open* **1**, e181018 (2018).

183. Zhang, H. *et al.* Regularized estimation in sparse high-dimensional multivariate regression, with application to a DNA methylation study. *Statistical Applications in Genetics and Molecular Biology* **16**, 159-171 (2017).

184. Rondina, J.M. *et al.* SCoRS--A Method Based on Stability for Feature Selection and Mapping inNeuroimaging [corrected]. *IEEE Transactions on Medical Imaging* **33**, 85-98 (2014).

185. Zhang, X. *et al.* Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* **7**, 197 (2006).

186. Darst, B.F., Malecki, K.C. & Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics* **19**, 65 (2018).

187. Pinaya, W.H. *et al.* Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports* **6**, 38897 (2016).

188. Malison, R.T. *et al.* Inter-rater reliability and concurrent validity of DSM-IV opioid dependence in a Hmong isolate using the Thai version of the Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA). *Addictive Behaviors* **36**, 156-160 (2011).

189. Mohamed Hebiri, J.C.L. How Correlations Influence Lasso Prediction. *arXiv* (2012).

190. Chih-wei Hsu , C.-c.C., Chih-jen Lin. A practical guide to support vector classification. Technical Report, Department of Computer Science and Information Engineering, University of National Taiwan, Taipei (2010). https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

191. Thais Mayumi Oshiro, P.S.P., José Augusto Baranauskas. How Many Trees in a Random Forest? *8th International Conference on Machine Learning and Data Mining in Pattern Recognition.* **July**(2012).

192. Debaditya Roy, S.R.M.K., Krishna Mohan Chalavadi. Feature selection using Deep Neural Networks. *2015 International Joint Conference on Neural Networks (IJCNN)* (2015).

193. Halligan, S., Altman, D.G. & Mallett, S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European Radiology* **25**, 932-939 (2015).

194. Mohamed Bekkar, H.K.D., Taklit Akrouf Alitouche & Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications* **3**(2013).

195. Weng, W.H., Wagholikar, K.B., McCray, A.T., Szolovits, P. & Chueh, H.C. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making* **17**, 155 (2017).

196. Saito, T. & Rehmsmeier, M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics* **33**, 145-147 (2017).

197. Fabian Pedregosa, G.V., Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

198. Gowin, J.L., Ball, T.M., Wittmann, M., Tapert, S.F. & Paulus, M.P. Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse. *Drug and Alcohol Dependence* **152**, 93-101 (2015).

199. Erguzel, T.T. *et al.* Binomial Logistic Regression and Artificial Neural Network Methods to Classify Opioid-Dependent Subjects and Control Group Using Quantitative EEG Power Measures. *EEG and Clinical Neuroscience Society*, 1550059418824450 (2019).

200. Ahn, W.Y. & Vassileva, J. Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. *Drug and Alcohol Dependence* **161**, 247-257 (2016).

201. Ahn, W.Y., Ramesh, D., Moeller, F.G. & Vassileva, J. Utility of Machine-Learning Approaches to Identify Behavioral Markers for Substance Use Disorders: Impulsivity Dimensions as Predictors of Current Cocaine Dependence. *Frontiers in Psychiatry* **7**, 34 (2016).

202. Buu, A., Johnson, N.J., Li, R. & Tan, X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine* **30**, 2326-40 (2011).

203. Robins, L.N. The intimate connection between antisocial personality and substance abuse. *Social Psychiatry and Psychiatric Epidemiology* **33**, 393-9 (1998).

204. Brook, J.S., Whiteman, M., Finch, S.J. & Cohen, P. Young adult drug use and delinquency: childhood antecedents and adolescent mediators. *Journal of the American Academy of Child and Adolescent Psychiatry* **35**, 1584-1592 (1996).

205. Dragisic, T., Dickov, A., Dickov, V. & Mijatovic, V. Drug Addiction as Risk for Suicide Attempts. *Materia Socio Medica* **27**, 188-191 (2015).

206. Phillips, J., Carpenter, K.M. & Nunes, E.V. Suicide risk in depressed methadone-maintained patients: associations with clinical and demographic characteristics. *American Journal of Addiction* **13**, 327-32 (2004).

207. Harned, M.S., Najavits, L.M. & Weiss, R.D. Self-harm and suicidal behavior in women with comorbid PTSD and substance dependence. *The American Journal on Addictions* **15**, 392-395 (2006).

208. Tremeau, F. *et al.* Suicidality in opioid-dependent subjects. *The American Journal on Addictions* **17**, 187-194 (2008).

209. Oliveto, A.H., Feingold, A., Schottenfeld, R., Jatlow, P. & Kosten, T.R. Desipramine in opioid-dependent cocaine abusers maintained on buprenorphine vs methadone. *JAMA Psychiatry* **56**, 812-820 (1999).

210. Krantz, M.J. & Mehler, P.S. Treating opioid dependence. Growing implications for primary care. *JAMA Internal Medicine* **164**, 277-288 (2004).

211. Schottenfeld, R.S., Pakes, J., Ziedonis, D. & Kosten, T.R. Buprenorphine: dose-related effects on cocaine and opioid use in cocaine-abusing opioid-dependent humans. *Biological Psychiatry* **34**, 66-74 (1993).

212. Mannelli, P., Wu, L.T., Peindl, K.S. & Gorelick, D.A. Smoking and opioid detoxification: behavioral changes and response to treatment. *Nicotine & Tobacco Research* **15**, 1705-1713 (2013).

213. Thylstrup, B., Schroder, S. & Hesse, M. Psycho-education for substance use and antisocial personality disorder: a randomized trial. *BMC Psychiatry* **15**, 283 (2015).

214. Carpenter, C.S. Workplace drug testing and worker drug use. *Health Services Research* **42**, 795-810 (2007).

215. Barry, D.T., Steinberg, M.A., Wu, R. & Potenza, M.N. Characteristics of black and white callers to a gambling helpline. *Psychiatric Services* **59**, 1347-1350 (2008).

216. Welte, J.W., Barnes, G.M., Tidwell, M.C., Hoffman, J.H. & Wieczorek, W.F. Gambling and Problem Gambling in the United States: Changes Between 1999 and 2013. *Journal of Gambling Studies* **31**, 695-715 (2015).

217. Rounsaville, B.J. & Kleber, H.D. Untreated opiate addicts. How do they differ from those seeking treatment? *Archives of General Psychiatry* **42**, 1072-1077 (1985).

218. Johnson, R.E. *et al.* A comparison of levomethadyl acetate, buprenorphine, and methadone for opioid dependence. *New England Journal of Medicine* **343**, 1290-1297 (2000).

219. Fiellin, D.A. *et al.* Drug treatment outcomes among HIV-infected opioid-dependent patients receiving buprenorphine/naloxone. *Journal of Acquired Immune Deficiency Syndromes* **56 Suppl 1**, S33-38 (2011).

220. Kazakova, O.B. *et al.* [Synthesis and pharmacological activity of amides and ozonolysis product of maleopimaric acid]. *Bioorganicheskaya Khimiya* **36**, 832-840 (2010).

221. Liang, W. & Chikritzhs, T. Reduction in alcohol consumption and health status. *Addiction* **106**, 75-81 (2011).

222. Wells, E.A. *et al.* Race differences in antisocial behaviors and attitudes and early initiation of substance use. *Journal of Drug Education* **22**, 115-130 (1992).

223. MacDonald, G.J. Atheist gatherings are mostly white, male. *Religion News Service* (2011).

224. Roberts, A.L., Gilman, S.E., Breslau, J., Breslau, N. & Koenen, K.C. Race/ethnic differences in exposure to traumatic events, development of post-traumatic stress disorder, and treatment-seeking for post-traumatic stress disorder in the United States. *Psychological Medicine* **41**, 71-83 (2011).

225. Moscati, A. & Mezuk, B. Losing faith and finding religion: religiosity over the life course and substance use and abuse. *Drug and Alcohol Dependence* **136**, 127-134 (2014).

226. Smyth, B.P., Barry, J., Keenan, E. & Ducray, K. Lapse and relapse following inpatient treatment of opiate dependence. *Irish Medical Journal* **103**, 176-179 (2010).

227. Scherbaum, N. & Specka, M. Factors influencing the course of opiate addiction. *International Journal of Methods in Psychiatric Research* **17 Suppl 1**, S39-44 (2008).

228. Junghwan Cho, K.L., Ellie Shin, Garry Choy, Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *ICLR2016* (2016).

229. Shiraly, R. & Taghva, M. Factors Associated with Sustained Remission among Chronic Opioid Users. *Addiction & Health* **10**, 86-94 (2018).

230. Agrawal, A. *et al.* The genetics of addiction-a translational perspective. *Translational Psychiatry* **2**, e140 (2012).

231. Bevilacqua, L. & Goldman, D. Genes and addictions. *Clinical Pharmacology & Therapeutics* **85**, 359-361 (2009).

232. Charbogne, P., Kieffer, B.L. & Befort, K. 15 years of genetic approaches in vivo for addiction research: Opioid receptor and peptide gene knockout in mouse models of drug abuse. *Neuropharmacology* **76 Pt B**, 204-217 (2014).

233. Finno, C.J., Aleman, M., Higgins, R.J., Madigan, J.E. & Bannasch, D.L. Risk of false positive genetic associations in complex traits with underlying population structure: a case study. *Journal of Veterinary Science* **202**, 543-549 (2014).

234. Kitsios, G.D., Tangri, N., Castaldi, P.J. & Ioannidis, J.P. Laboratory mouse models for the human genome-wide associations. *PLoS One* **5**, e13782 (2010).

235. Palmer, A.A. & de Wit, H. Translational genetic approaches to substance use disorders: bridging the gap between mice and humans. *Human Genetics* **131**, 931-939 (2012).

236. Guenet, J.L. The mouse genome. *Genome Research* **15**, 1729-1740 (2005).

237. Marchant, N.J., Li, X. & Shaham, Y. Recent developments in animal models of drug relapse. *Current Opinion in Neurobiology* **23**, 675-683 (2013).

238. Fujii, K., Koshidaka, Y., Adachi, M. & Takao, K. Effects of chronic fentanyl administration on behavioral characteristics of mice. *Neuropsychopharmacology Reports* **39**, 17-35 (2019).

239. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

240. Miller, J.A., Horvath, S. & Geschwind, D.H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12698-12703 (2010).

241. Scarpa, J.R. *et al.* Cross-species systems analysis identifies gene networks differentially altered by sleep loss and depression. *Science Advances* **4**, eaat1294 (2018).

242. Baker, E.J., Jay, J.J., Bubier, J.A., Langston, M.A. & Chesler, E.J. GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Research* **40**, D1067-76 (2012).

243. Goldberg, L.R. Systems Genetics Analysis of Addiction-Associated Traits. *Boston University School of Medicine dissertation* (2017). https://open.bu.edu/handle/2144/23561

244. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Research* **7**, 1860 (2018).

245. Ritchie, M.E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

246. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).

247. Durinck, S., Spellman, P.T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 1184-1191 (2009).

248. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90-97 (2016).

249. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinburgh)* **95**, 221-227 (2005).

250. Vorontsov, I.E., Kulakovskiy, I.V. & Makeev, V.J. Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for Molecular Biology* **8**, 23 (2013).

251. Kristin M. Mignogna, S.A.B., Brien P. Riley, Aaron R. Wolen , Michael F. Miles Cross-species analysis reveals alcohol dependence-associated gene networks. *bioRxiv preprint* (2018).

252. Huelga, S.C. *et al.* Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Reports* **1**, 167-78 (2012).

253. Ruan, Q.T. *et al.* A mutation in Hnrnph1 that decreases methamphetamine-induced reinforcement, reward, and dopamine release and increases synaptosomal hnRNP H and mitochondrial proteins. *Journal of Neuroscience* (2019).

254. Xu, J. *et al.* A heroin addiction severity-associated intronic single nucleotide polymorphism modulates alternative pre-mRNA splicing of the mu opioid receptor gene OPRM1 via hnRNPH interactions. *Journal of Neuroscience* **34**, 11048-11066 (2014).

255. Yazdani, N. *et al.* Hnrnph1 Is A Quantitative Trait Gene for Methamphetamine Sensitivity. *PLoS Genetics* **11**, e1005713 (2015).

256. Dent, C.L. *et al.* Impulsive Choice in Mice Lacking Paternal Expression of Grb10 Suggests Intragenomic Conflict in Behavior. *Genetics* **209**, 233-239 (2018).

257. Johnson, C. *et al.* Genome wide association for substance dependence: convergent results from epidemiologic and research volunteer samples. *BMC Medical Genetics* **9**, 113 (2008).

258. Logue, J.S. & Morrison, D.K. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. *Genes & Development* **26**, 641-650 (2012).

259. Yan, X. *et al.* Deletion of the Imprinted Gene Grb10 Promotes Hematopoietic Stem Cell Self-Renewal and Regeneration. *Cell Reports* **17**, 1584-1594 (2016).

260. Himburg, H.A. *et al.* Dickkopf-1 promotes hematopoietic regeneration via direct and niche-mediated mechanisms. *Nature Medicine* **23**, 91-99 (2017).

261. Lisse, T.S. & Hewison, M. Vitamin D: a new player in the world of mTOR signaling. *Cell Cycle* **10**, 1888-1889 (2011).

262. Aranow, C. Vitamin D and the immune system. *Journal of Investigative Medicine* **59**, 881-886 (2011).

263. Pasala, S., Barr, T. & Messaoudi, I. Impact of Alcohol Abuse on the Adaptive Immune System. *Alcohol Research* **37**, 185-197 (2015).

264. Liang, X., Liu, R., Chen, C., Ji, F. & Li, T. Opioid System Modulates the Immune Function: A Review. *Translational Perioperative and Pain Medicine* **1**, 5-13 (2016).

265. Hanson, R.W. Glycolysis and Gluconeogenesis *Biomedical education* **9**(1981).

266. Mysels, D.J. & Sullivan, M.A. The relationship between opioid and sugar intake: review of evidence and clinical applications. *Journal of Opioid Management* **6**, 445-452 (2010).

267. Najafipour, H. & Beik, A. The Impact of Opium Consumption on Blood Glucose, Serum Lipids and Blood Pressure, and Related Mechanisms. *Frontiers in Physiology* **7**, 436 (2016).

268. Nies, V.J. *et al.* Fibroblast Growth Factor Signaling in Metabolic Regulation. *Frontiers in Endocrinology* **6**, 193 (2015).

269. Goldstein, R.Z. & Volkow, N.D. Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nature Reviews. Neuroscience* **12**, 652-669 (2011).

270. Jerome Friedman, T.H., Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1-22 (2010).

271. Krstajic, D., Buturovic, L.J., Leahy, D.E. & Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* **6**, 10 (2014).

272. Gilles Louppe, L.W., Antonio Sutera and Pierre Geurts. Understanding variable importances in forests of randomized trees. *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems* **1**, 431-439 (2013).

273. Breiman, L. Manual on setting up, using, and understanding random forests v3. 1. *Random forests. Machine Learning,* **45**, 5-32 (2001).

274. Chollet, F. keras. *Github repository* (2015).

275. Diederik P. Kingma, J.L.B. Adam: a method for stochastic optimization. *ICLR conference* (2015).

276. Matthieu Courbariaux, Y.B., Jean-Pierre David. BinaryConnect: Training Deep Neural Networks with binary weights during propagations.

277. Adam Paszke, A.C., Sangpil Kim, Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* (2016).

278. Chintala, A.R.L.M.S. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR conference* (2016).

279. Tieleman, T.a.H., G. Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning. *Technical report* (2012).

280. Duchi, J., Hazan, Elad, and Singer, Yoram. . Adaptive subgradient methods for online learning and stochastic optimization. . *The Journal of Machine Learning Research* **12**, 2121–2159 (2011).

281. Li, Y., Chen, C.Y. & Wasserman, W.W. Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. *Journal of Computational Biology* **23**, 322-36 (2016).

282. Chun-Hao Chang, L.R., Anna Goldenberg. Dropout Feature Ranking for Deep Learning Models. arXiv:1712.08645(2018).

283. Kirkpatrick, S.L. & Bryant, C.D. Behavioral architecture of opioid reward and aversion in C57BL/6 substrains. *Frontiers in Behavioral Neuroscience* **8**, 450 (2015).

284. Bryant, C.D. *et al.* A role for casein kinase 1 epsilon in the locomotor stimulant response to methamphetamine. *Psychopharmacology* **203**, 703-711 (2009).

285. Bryant, C.D. *et al.* Csnk1e is a genetic regulator of sensitivity to psychostimulants and opioids. *Neuropsychopharmacology* **37**, 1026-1035 (2012).

286. Yazdani, N., Shen, Y., Johnson, W.E. & Bryant, C.D. Striatal transcriptome analysis of a congenic mouse line (chromosome 11: 50-60Mb) exhibiting reduced methamphetamine sensitivity. *Genomics Data* **8**, 77-80 (2016).

287. Liao, Y., Smyth, G.K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41**, e108 (2013).

288. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

CURRICULUM VITAE