Title
Integration of evidence across human and model organism studies: A meeting report

Journal
Genes Brain & Behavior, 20(6)

ISSN
1601-1848

Authors
Palmer, Rohan HC
Johnson, Emma C
Won, Hyejung
et al.

Publication Date
2021-07-01

DOI
10.1111/gbb.12738

Peer reviewed

Title: Integration of Evidence across Human and Model Organism Studies: A Meeting Report

Rohan H. C. Palmer[1*], Emma C. Johnson[2], Hyejung Won[3], Renato Polimanti[4], Manav Kapoor[5], Apurva Chitre[6], Molly A. Bogue[7], Chelsie E. Benca-Bachman[1], Clarissa C. Parker[8], Oana Ursu[9], Anurag Verma[10], Timothy Reynolds[7], Jason Ernst[11], Michael Bray[2], Soo Bin Kwon[11], Dongbing Lai[12], Bryan C. Quach[13], Nathan C. Gaddis[13], Laura Saba[14], Hao Chen[15], Michael Hawrylycz[16], Shan Zhang[17], Yuan Zhou[18], Spencer Mahaffey[19], Christian Fischer[20], Sandra Sanchez-Roige[6], Anita Bandrowski[21], Qing Lu[18], Li Shen[22], Vivek Philip[7], Joel Gelernter[4], Laura J. Bierut[2], Dana B. Hancock[13], Howard J. Edenberg[12,23], Eric O. Johnson[13], Eric J. Nestler[22], Peter B. Barr[24], Pjotr Prins[25], Desmond J. Smith[26], Schahram Akbarian[27], Thorgeir Thorgeirsson[28], Dave Walton[7], Erich Baker[29], Daniel Jacobson[30,31], Abraham A. Palmer[6,32], Michael Miles[33], Elissa J. Chesler[7], Jake Emerson[7], Arpana Agrawal[2], Maryann Martone[20], Robert W. Williams[20]

[1] Behavioral Genetics of Addiction Laboratory, Department of Psychology, Emory University, Atlanta, GA, USA

[2] Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA

[3] Department of Genetics and Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC (only USA, Canada and Australia), USA

[4] Department of Psychiatry, Yale University School of Medicine, West Haven, CT, USA

[5] Nash Family Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029 USA

[6] Department of Psychiatry, University of California, San Diego, CA, USA

[7] The Jackson Laboratory, Bar Harbor, ME, USA

[8] Department of Psychology and Program in Neuroscience, Middlebury College, Middlebury, VT, USA

[9] The BROAD Institute, Massachusetts Institute of Technology, Cambridge, MA, USA

[10] Biomedical and Translational Informatics Laboratory, University of Pennsylvania, Philadelphia, PA, USA

[11] Department of Biological Chemistry, University of California, Los Angeles, Los Angeles, CA, US

[12] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, US

[13] GenOmics, Bioinformatics, and Translational Research Center, Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA

[14] Department of Pharmaceutical Sciences, University of Colorado, Anschutz Medical Campus, Aurora, CO

[15] Department of Pharmacology, Addiction Science, and Toxicology, University of Tennessee Health Science Center, Memphis, TN, USA

[16]Allen Institute, Seattle, United States

[17]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

[18]Department of Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA

[19] Department of Pharmaceutical Sciences, School of Pharmacy, University of Colorado Denver, Aurora, CO 80045, USA

[20] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

[21] Department of Neuroscience, UCSD, SciCrunch Inc, La Jolla, CA 92093, USA.

[22] Icahn School of Medicine at Mount Sinai, New York, New York

[23] Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN

[24] Department of Psychology, Virginia Commonwealth University, Richmond, VA, USA

[25] Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA

[26] Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA

[27] Friedman Brain Institute and Departments of Psychiatry and Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY USA

[28] deCODE Genetics/Amgen Inc., Sturlugata 8, 101 Reykjavik, Iceland

[29] Department of Computer Science, Baylor University, Waco, TX, USA

[30] Computational and Predictive Biology, Biosciences, Oak Ridge National Laboratory, Oak Ridge, TN (only USA, Canada, and Australia), USA

[31] Department of Psychology, University of Tennessee Knoxville, Knoxville, TN, USA

[32] Institute for Genomic Medicine, University of California San Diego, La Jolla, CA 92093.

[33] Department of Pharmacology and Toxicology, Virginia Commonwealth University, Richmond, VA, USA

Corresponding Author:
Rohan H.C. Palmer
Department of Psychology
Emory University
36 Eagle Row, Atlanta GA, 30322
Rohan.Palmer@Emory.edu
Tel: +1 404-727-3126

Manuscript info:

# of words (abstract): 177

# of words (cover + references): 6455

# of tables: 1

# of figures: 0

Keywords: Drug Abuse, Working Group, Genomics, Model Organisms, Multi-omic

Abstract

The National Institute on Drug Abuse and Joint Institute for Biological Sciences at the Oak Ridge National Laboratory hosted a meeting attended by a diverse group of scientists with expertise in substance use disorders (SUDs), computational biology, and FAIR (Findability, Accessibility, Interoperability, and Reusability) data sharing. The meeting's objective was to discuss and evaluate better strategies to integrate genetic, epigenetic, and 'omics data across human and model organisms to achieve deeper mechanistic insight into SUDs. Specific topics were to (a) evaluate the current state of substance use genetics and genomics research and fundamental gaps, (b) identify opportunities and challenges of integration and sharing across species and data types, (c) identify current tools and resources for integration of genetic, epigenetic, and phenotypic data, (d) discuss steps and impediment related to data integration, and (e) outline future steps to support more effective collaboration—particularly between animal model research communities and human genetics and clinical research teams. This review summarizes key facets of this catalytic discussion with a focus on new opportunities and gaps in resources and knowledge on SUDs.

Keywords: GWAS; data integration; cross-species; substance use disorders

1. Introduction

On May 29–31, 2019, the National Institute on Drug Abuse (NIDA) and the Joint Institute for Biological Sciences at the Oak Ridge National Laboratory (ORNL) hosted the *Addiction Genetics and Epigenetics Data Jamboree* meeting at Oak Ridge, Tennessee. Over thirty scientists with expertise in genetics and genomics of substance use in human and model organisms gathered to discuss linking data and results across systems that exploit genetics, genomics, epigenetics, and other omics by leveraging innovative statistical methods and computational tools. The meeting commenced with an open discussion of the state of substance use genetics, including the strengths and weaknesses of various approaches to genotype-phenotype associations in humans and model organisms. Most notably, researchers discussed how joint data- and theory-driven studies using integrative cross-species and multi-omics approaches could more rapidly discover and translate mechanisms than relying upon genome-wide association studies (GWAS) or model organisms alone. Over the course of two days, researchers participated in thematic discussions that centered on the current state of knowledge, gaps in understanding and advantages and challenges of: (1) data analyses using multi-species and multi-omic data, (2) data integration methods/procedures, and (3) multi-omic data generation and sharing/accessibility. Meeting participants reconvened on the third day to summarize findings and since then have reflected upon the field's latest findings around the meeting's topical areas in the preparation of the current document. Each researcher brought their unique experience, perspective, and expertise to these discussions, and a consensus was not always reached for the best path forward on every topic. Not all authors of this report necessarily endorse all ideas presented herein.

This report aims to summarize the discussions by focusing on the state of science, including opportunities for more effective cross-talk and collaboration between human and model organism research communities, as well as barriers to data acquisition and integration. Next, we discuss the methods and tools used for genetic and genomic discovery, their assumptions and limitations, as well as areas for improvement needed to achieve rapid translation of genetic loci to identified mechanisms and potential treatments. We review challenges of data transportability and sharing (i.e., Findability, Accessibility, Interoperability,

and Reusability data practices), for which there are interpersonal, legal, and technological barriers of integrating diverse data types. Finally, we describe some gaps to address in future programs on substance use disorders (SUDs).

Status of Substance Use and Disorders Genetics and Genomics

SUDs represent a pressing area of unmet medical, psychological, and social needs. In 2017, alcohol and illicit substance use and disorders resulted in 13,969 and 67,000 deaths (directly and indirectly) in the United States, respectively,[1] which was less than smoking (~250,000 deaths), but more than liver disease (62,493 deaths)[2] and diabetes (68,558 deaths).[3] Worldwide, SUDs have a relatively early onset and contribute to approximately 21% of lost disability-adjusted life years[4] (15% for smoking and second-hand smoke not counting comorbid drug use[1]), emphasizing the high societal and personal cost to affected individuals and communities. Twin- and family-based studies show that SUDs generally have moderate to high heritability,[5] with sequence differences contributing to 50–70% of variance in liability. Large-scale GWASs investigating hundreds of thousands of participants have become a reliable method to localize and identify genomic regions, genes, and common and substance-specific nucleotide differences that contribute to the heritability of the many facets of SUDs. [6-8]

To date, there has been substantial progress in the characterization of the genetic etiology of human SUDs.[9-13] Data sharing, meta-analysis, and very large sample sizes have begun to yield loci for alcohol-,[14-19] tobacco-,[18,20] and cannabis-related traits.[20] The past three years have witnessed an escalation in these discoveries - for instance, findings for alcohol use disorder (AUD) increased from one locus (N=14,904 cases) in 2018 to 29 independent variants in 2020 (N=435,563, including >57,000 cases). These human GWASs have shown that SUDs are highly polygenic. This polygenicity may be partially explained by human-specific evolutionary pressures and diagnostic heterogeneity.[21] Notably, the history of SUD and psychiatric GWAS has shown that more common variants with modest effect sizes can be identified and replicated when studies are well-powered. Yet, there are other substances of abuse for which we still lack sufficient power (e.g., opioids[22] and cocaine[23]) for unbiased identification of the heritable

components of susceptibility, severity, and relapse. For most common diseases, the number of genome-wide significant hits that are discovered increases sharply after a threshold sample size that ranges from about 10,000 to 100,000.[24] In the case of psychiatric disease, it took 36,989 cases and 113,075 controls to identify 108 loci for schizophrenia.[25] A simulation study by Walters et al. 2019 suggested that AUD and other related SUDs[26] have effect size distributions similar to major depression,[27] a disease that required approximately 10,000 cases to identify the first locus,[28] and may require sample sizes between 55,000 and 130,000 cases (or more) to identify large numbers of commonly occurring variants. [15] While biobanks and electronic health records provide opportunities for increasing sample sizes for AUD, the ability to adequately assess illicit drug use disorder from biobanks remains questionable. That said, steady progress is being made for illicit substances. For example, a recently published GWAS for opioid use disorder (OUD) in the Million Veterans Program and two additional samples, obtained genome-wide significance for rs1799971 in the gene encoding the mu-opioid receptor, OPRM1, with 8,529 cases and 71,200 opioid-exposed controls[22] though additional work is needed to validate these findings.

It is also important to note that identifying genetically-mediated mechanisms of disease is also partially contingent on how well a phenotype is defined so that it reflects relevant biological and environmental variation. In human GWAS, phenotypic heterogeneity, which is evident in diagnostic classification, as well as the imprecision of recall and self-report, has been shown to result in low heritability (in some instances) and specificity for disease prediction.[29] Compared to humans, model organisms have the advantages of narrowly defined phenotypic assays applied to both experimental and control groups and objective measurements. However, animal models poorly reflect the interpersonal and quality of life aspects of human SUD.[30] Human studies using case-control and quantitative phenotypes of the most predominantly used substances, alcohol and tobacco, with sufficiently large sample size have recently confirmed suspected genetic mediation of pharmacokinetic and pharmacodynamic pathways; studies also suggest greater relevance of single nucleotide variants expressed in brain [31-33]. Liu et al. 2019[18] found that all central-nervous-system-expressed nicotinic receptor genes (except for *CHRNA7*) were significantly associated with one or more smoking phenotypes that they examined. This suggests that related phenotypes,

such as age of smoking initiation and cigarettes per day, may show overlapping but differential patterns of associations with relevant genetic variation. Therefore, it is important to examine a variety of different phenotypes, from case-control phenotypes to endophenotypes. For example, in a GWAS of a pharmacologically relevant phenotype for smoking, a measure of the rate of nicotine metabolism (the nicotine metabolite ratio [NMR]), identified polymorphisms that accounts for nearly 40% of the phenotypic variance in NMR,[34] but these same loci do not have a similarly large effect on nicotine dependence. Consequently, there is still a gap in understanding the broad and substance-specific mechanisms and the functional significance of DNA variants that have been discerned to date using endo-, clinical-, and coarse phenotypes and biomarkers. Some researchers at the meeting commented that mixed-linear-model-based and traditional GWAS and quantitative trait locus (QTL) analyses alone cannot solve these phenotype limitations because the variance structure of agglomerative phenotypes does not match that of the genome and the associated structures/tissues. Others countered that well-powered GWAS complemented by new post-hoc computational methods (e.g., genomic structural equation modeling [35] and multivariate GWAS [36], to name a few) might surmount minimal phenotyping limitations. For a detailed example of deep phenotyping issues in a complex psychiatric disorder, we recommend the recent paper by Cai et al. 2020.[29]

Based on these observations, researchers recognized that other methods should help complement and extend well-powered GWAS methods to address current knowledge gaps in the genetic architecture of SUDs. A notable illustration arises from the characterization of the complement C4 pathway in schizophrenia, which arose from a GWAS that identified a strong signal in the MHC locus but required deep, cross-species cellular and molecular experiments to explicate. Previous studies[15,37] have also indicated this will require (1) larger sample sizes, (2) better phenotyping, (3) more diverse samples, (4) improved coverage of genetic variation by GWAS arrays or greater emphasis on sequencing,[38-40] and (5) more comprehensive system-based models and hypotheses that incorporate epistasis (GxG), environmental factors, GxE, and many comorbidities. Systems-based and multi-level studies would ideally model the complex nature of SUDs using multiple cofactors (and confounders) and take into account the inevitability that many agglomerative phenotypes will be made up of multiple mechanistically distinct sub-

phenotypes. In addition to the more nuanced and precisely defined and quantified phenotypes and cofactors (e.g., BMI for alcohol[41]) and confounders,[42] such studies would also incorporate other forms of DNA variation and potential non-linear (i.e., GxG and GxE) effects - although recent studies have suggested that most of the genetic variance for complex traits appears to be largely due to additive effects, with negligible dominance effects, and an indeterminate amount of epistatic effects due to power and study design issues.[43] While the importance of these different issues and approaches was discussed, a diversity of opinions was expressed about GxG effects, and the group did not reach a consensus. Still, it is worth noting that a negligible genome-wide contribution of dominance effects does not preclude the existence of individual loci with a dominant mode of inheritance. While the importance of these different issues and approaches was discussed, a diversity of opinions was expressed about GxG effects, and the group did not reach consensus.

At the sequence level, many studies are also still missing significant genetic diversity—particularly from non-European populations.[44] Even though copy number variant (CNV) studies of psychiatric disorders are becoming more commonplace,[45] mobile element polymorphisms, inversions and other types of structural variants are still missed in GWAS—as are subsets of variants not tagged using standard GWAS arrays or incorrectly aligned to a single canonical reference genome. In short, recent insights from past studies highlight how gaps in our understanding could be addressed using large and genetically diverse samples (is being achieved for nicotine and alcohol, but not other substances), better phenotyping, new computational methods, and long-read sequencing technologies to capture and model causal genome variants, especially those (e.g., CNVs, insertions, deletions, and inversions) not well captured by GWAS arrays; see Peterson et al. 2019 [46] for a detailed discussion on opportunities for diversity in GWAS. In addition, single-cell technologies, such as single-cell-RNA-seq, and complementary approaches towards studying regulatory effects of variants, among others, will help to better uncover cell-type specific networks involved in SUDs, as has been documented for schizophrenia.[47] Altogether, these types of systems-based approaches that incorporate multiple layers of genomic and environmental data will require advanced methods, that may include multilevel machine learning, deep learning, and explainable-artificial intelligence techniques to

name a few; and these model-free approaches will have to accommodate features specific to the human genome, such as population substructure, which can confound association signals.[48] Likewise, it will require a more comprehensive, integrated capture of population-scale data at multiple omics layers (genome, epigenome, transcriptome, metabolome, microbiome) in both model organism and human studies (see Table 1). Costs for generating multi-omic data, including brain proteomics and metabolomics are falling rapidly and making such programs possible.

Complementary to human GWAS, research using model organisms is amassing a large body of evidence supporting causal roles for many genomic loci and gene variants related to SUDs (e.g., *Taar1* for methamphetamine'[49] *APBA2* for addiction,[46] *XRCC5* for alcohol dependence,[50] and the use of CRISPy Critters for instance in alcohol research[51]). Still, these findings probe only a small part of the complex central nervous system (CNS) molecular and cellular networks affected by addictive substances. There is also deep sequence data on shorter classes of DNA variants and expression data collected in many contexts across large populations of key model organisms, including *Drosophila* (the *Drosophila* Genetic Reference Panel),[52,53] mouse (Collaborative Cross, the Hybrid Mouse Diversity panel, and the BXD family, collectively $n > 200$ isogenic strains,[54,55] and outbred mouse populations, including several heterogeneous stocks,[56-59] advanced intercross lines[60]), and rat populations (e.g., Hybrid Rat Diversity Panel and the National Institute of Health (NIH) heterogeneous stock[60], and outbred Sprague Dawley[61,62]). As a field, behavior geneticists, both human and animal modelers, are beginning to catalog and even understand the function(s) of subsets of variants that alter protein-coding sequence, modulate transcript and protein isoforms, or change expression.[63-65] However, although great progress has been made, we highlight key gaps:

1. the comparative invisibility of mobile element polymorphisms, some types of structural variants, simple tandem repeats, and rare variants, including *de novo* mutations;
2. the problematic nature of aligning a sequence to a linear reference genome rather than to pangenomes that are savvy with respect to sequence differences among individuals and ancestries; and

3. the reliance on simple additive models that cannot detect or are confounded by gene-by-gene epistatic interactions or cleanly dissect and unconfound GxE effects.[64,66]

Researchers at the meeting discussed gaps in knowledge and possibilities for the next phase of functional discovery for substance use and disorders, which will likely require (1) the construction of appropriate resources for systematic evaluation of loci function in humans, (2) quantitative experimental studies of SUDs in model organisms with a more realistic level of genetic complexity, (3) concerted multidisciplinary efforts to acquire additional samples for discovery/validation, and (4) a shift towards causal models and quasi-experimental research designs in order to understand gene-by-environment, gene-by-development, and epigenetic modifiers across a range of genetically-admixed and genetically simple cohorts of model organisms.

Theme A: Bridging the Gap between Human and Animal Research

Prioritizing variants for functional follow-up

In recent years, larger human GWAS have begun to produce a more robust and reliable set of genomic loci and gene variants. Similarly, model system studies complement these phenotype-genotype associations via behavioral neurogenetic methods, but not without limitations (see Table 1). Indeed, human and model organism studies offer varying degrees of power and limitations to identify a gene or network for functional follow-up. For example, human GWAS require very large samples to study phenotypes that may be less proximal to the biological elements. Model organisms require smaller sample sizes, but their individual single nucleotide polymorphisms (SNPs) and genes may not entirely map onto human biology and the substance use phenotypes that operate in a complex, human environment. Given that the collection of larger, more diverse GWAS samples for SUD phenotypes will require targeted data collection, especially in underrepresented populations, some researchers at the meeting acknowledged that animal QTL, and other methods (e.g., recombinant inbred strains[55]), can help make headway in parallel. One area for further development includes refinement of efficient and unbiased computational workflows to rank top variants and map their target genes and gene,

molecular, and cellular networks.

Researchers at the meeting discussed strategies to make advances in using integrative approaches, which could rapidly locate and translate loci for SUDs. These strategies combine data from GWAS in humans with well-matched experimental work in model organisms—both genetically admixed crosses and gene knockout and knock-in studies. Ideally, these studies would leverage a universal platform for sharing current datasets from model organisms with human GWAS findings, a resource currently lacking. At the time of this publication, data from model organism studies are largely isolated by species and even by strain and type. As such, they are often far from FAIR compliant[67] and are just as hard to access and integrate as GWAS data from heterogeneous human populations, which are not all shared on the NIH's database of Genotypes and Phenotypes (dbGaP) or other repositories available to the scientific community. These realities further compound the challenge of rigorously combining human and animal model data sets (see Theme C discussion for details).

Why data integration across species and multiple omics is important for expansion, discovery, and translation of genetic risk for SUDs

While there are many differences between behaviors, body, and brain structures of all model organisms and humans, there is still a high level of genomic and functional commonality that can be leveraged under tightly controlled environmental and treatment conditions. In essence, a randomized controlled trial across multiple genotypes can usually be designed and implemented reasonably easily with model organisms.[68] Likewise, causal models can be constructed to evaluate potential confounders by, for instance, comparing behavioral assays across constructed genetic backgrounds of varying disease susceptibility (see Table 1: Areas of Convergence). Molecular and cellular endophenotypes of SUDs are readily accessible in many model organisms. Conservation of functional genes and networks across species can provide genuine insight of high translational relevance–particularly when the GWAS searchlight has illuminated a small number of plausible genes and genomic regions. Because of differing evolutionary histories, individual variants among humans and model organisms are often not

conserved[69,70]; however, the prospects of comparing genetically engineered lines to diverse populations of mice holds significant promise for disease mapping and detecting epistatic interactions.[55] This apparent gap in the literature highlighted why analyses are best suited to be conducted at the level of genes, molecular networks, and gene sets. Still, attendees at the meeting acknowledged that experimental models could complement these analyses by providing a reproducible resource to identify fundamental processes and modifiers that affect aspects of SUD with the goal to transition as efficiently as possible to well-reasoned interventions that reduce SUD burden. Gene network perturbations that are evident in certain model organism experiments and humans may highlight novel entry points for pharmaceutical intervention and innovation that would be missed by the study of humans alone (e.g., modulation of an associated protein if variants are in a regulatory region). Further, identification of molecular and cellular networks that contribute to SUD risk, progression, and relapse will benefit from access to longitudinally collected datasets to strengthen causal inferences, define and test plausible models, and refine treatment options on the basis of genotypes and diplotypes.

Human tissues, cells, and organoids are highly useful tools for elucidating molecular and cellular networks in human-relevant model systems but have fundamental limitations, especially with respect to higher-order behavioral outcome variables that replicate aspects of human addiction. While formal proof of the roles of DNA variants is most readily provided using gene-engineered animals or specific pharmacological treatments, it is vital to note that "necessary and sufficient" causal criteria depend greatly on the genomic background [71]. Moreover, gene-engineered models will ideally account for genetic diversity in order to ensure that results are not only replicable but are likely to have external validity across species. While some researchers predicted that data generated from these approaches would show greater consilience with the diversity of human behavioral outcomes, others contended that additional research is needed to understand which animal paradigms and tissues best characterize the basic behavioral properties and neurobiological components of addiction, respectively.

Many researchers have begun to tackle the issue of variant prioritization by integrating multiple sources of information.[72-74] Indeed, most GWAS include detailed post-hoc analyses towards the identification of credible causal variants. Network integration is one method that can

permit the full illumination of patterns that are shared across gene sets derived from single omics data (e.g., genetic variants, RNA-seq in bulk tissue, single-cell RNA-seq, chromatin immunoprecipitation sequencing [ChIP-seq], ATAC-seq, methylome, etc.). Variant-based networks can be mapped onto genes, enabling a common basis for network integration: the gene level. A range of public data (e.g., ChIP-seq from ENCODE, RNA-seq from the Genotype-Tissue Expression [GTEx] project [75], Hi-C data for chromatin structure [76], protein-protein interaction data, etc.) can be incorporated to add evidence for the networks' biological plausibility; however several researchers advised caution as data limitations and improper handling could create biased results. Further sophisticated network layers can be generated with the use of new explainable-AI tools that can find highly accurate linear and nonlinear multi-way associations within and across omics layers [77]; though, as shown in the case of machine learning using a candidate SNPs for opioid dependence, extreme care should be taken to account for social inequities that permeate research practices and could likely confound biological mechanisms under study.[78] After integrating the networks from the different data inputs based on gene IDs, lines-of-evidence (LOE) scoring[79] methods offer a way to establish links between the networks, with each link adding to the score for connecting layers. Explainable-AI approaches such as Iterative Random Forest- Leave One Out Prediction (iRF-LOOP) are able to find linear and linear expression relationships in expression datasets derived from population-scale RNA-seq datasets and are more accurate than traditional co-expression approaches.[77] These explainable-AI derived networks can be built from publicly available datasets (such as GTEx) to provide tissue-specific regulatory patterns. They can similarly be built of single-cell-RNA-seq datasets to provide cell-type-specific regulatory networks. Of course, they can also be built from novel experimental data from individuals who were addicted to opioids. These networks can be combined with networks derived from other data types to form a multiplex network. For example, an explainable-AI-derived RNA expression network associated with opioid addiction in the nucleus accumbens (NAc) may link to a genome-wide epistasis (GWES)-based network[80] and a NAc-specific network assembled from the GTEx, and may also connect through to a protein-protein interaction network and signaling cascade network all through common gene IDs. Subsequently, Random Walk with Restart (RWR) approaches, which use an advanced form of network-association that is not limited to exploring shortest paths or nearest neighbors, can

jointly examine these multiple heterogeneous multiplex networks while retaining the critical topological information present in each network.[81] By jointly integrating multiple heterogeneous data layers, one can score and rank candidate genes from GWAS and genome-wide epistasis study (GWES) analyses using RWR-based LOE algorithms. This can help to prioritize genes from GWAS/GWES results and to provide mechanistic context for the resulting filtered genes sets by way of subnetworks that include the links among members of the filtered gene set and links to genes highly connected to members of the gene set in the network. This context greatly enhances mechanistic interpretation and the creation of conceptual models that can be used to design validation experiments in human tissue or animal models. Because similar gene-based networks can also be generated from model organisms, they can also be integrated with human networks via ortholog projection in order to leverage information from multiple species.

Challenges and Knowledge Gaps in Cross-Species Research

There is heterogeneity in the behavioral phenotypes and paradigms across humans and model organisms, respectively, that needs to be considered when attempting to identify the biobehavioral processes underlying substance use and disorders. Clinical diagnoses of SUDs in humans are based on assessments of drug-seeking, physical dependence, and social disruption but often struggle to quantify each of these phenotypes (e.g., the problem of going from a polythetic diagnosis to understanding severity/impact of combinations of criteria on a person's life).[82] It is often the case that qualitative symptoms are employed, and several combinations of criterion endorsements (i.e., 2 or more of 11 DSM-5 symptoms) could result in a diagnosis. This diagnostic heterogeneity (i.e., different case subjects meeting the criteria for endorsing varying sets of symptoms) leads to challenges in genetic mapping[83-85] and alignment with unconditioned and conditioned quantitative traits used in animal models. In contrast, animal studies place a high emphasis on measuring quantity/frequency and physiological dependence. Studies of alcohol and cannabis use disorders have shown quantitative and qualitative differences between the genetics of consumption quantity and frequency and the genetics of the disorders (e.g., impaired functioning, physical dependence, disruption of social responsibilities). [86,87] Likewise, a geneset derived from tobacco exposure paradigms in rodents shows modest enrichment for the SNP-

heritability of human tobacco consumption.[88] Notably, inbred strain comparison/selective breeding studies have allowed scientists to examine the effects of genetic background on multiple related traits.[89] Differences in the phenotypes assessed in humans and rodents may therefore contribute to a partially disconnected approach to understanding risk rather than a fully integrated approach, thus requiring detailed studies of consilience across phenotypes and omic-phenotype associations. For example, even just within humans, recent studies suggest that the genetics of human alcohol consumption, particularly frequency of alcohol intake, is only partly related to the genetics of alcohol problems (e.g., impaired functioning, physical dependence, disruption of social responsibilities).[19] Likewise, a geneset derived from tobacco exposure paradigms in rodents shows modest enrichment for the SNP-heritability of human tobacco consumption.[88] Therefore, differences in phenotypes and their associated genetic architecture, whether within or across organisms, should be taken into consideration, and leveraged when possible. As mentioned above, there is tremendous potential to build integrated, cross-species multi-omics networks that can serve to unify and utilize data and extant knowledge from both humans and model organisms.

There are several knowledge gaps that, if addressed, would help inform whether genetic results for SUD phenotypes can be translated across species. These included understanding (1) the degree of concordance among model organism findings, as well as (2) the extent to which model organism evidence generalizes to humans, (3) the contextual implication of tissue, sex, and ancestry on these effects, and (4) how unifying phenotypic definitions across databases can enhance sample sizes and data integration. To date, several studies have shown enrichment of mouse and rat gene sets (i.e., those that are differentially expressed in the presence of cocaine) in the human brain transcriptome for cocaine use disorder [90], as well as human GWAS of tobacco/nicotine consumption.[88] Identifying convergent genetic mechanisms between humans and model organisms in SUDs is an exciting challenge but is (relatively) close at hand. Even more daunting challenges (and rewards) are presented by the ambitious goal of identifying neural pathways conserved between model organisms and humans for addiction and its associated constellation of complex behaviors. Clearly, the molecular and bioinformatics tools that emerge from tackling the first problem will be a starting point for attacking the second.

Theme B: Current tools for integration of genetic, epigenetic, and phenotypic data

Several tools (e.g., methods, software, databases) currently exist and are under active development to aid scientists in analyzing and integrating multiple types and streams of data from a wide variety of model organisms and diverse human populations. Here we highlight a few that facilitate multi-omics and cross-species research. For a more comprehensive list of tools please see the paper by Reynolds et al. 2021.[91]

Functional mapping and annotation of genetic associations (FUMA) was developed [92] to annotate, prioritize, visualize, and interpret GWAS results. The application integrates genome-wide summary statistics with functional information, such as expression-QTL (eQTL) and chromosomal interaction mapping in a tissue-specific manner to identify the most likely causal SNPs. FUMA uses 18 biological data repositories (e.g., GTEx) and tools to functionally annotate GWAS hits. FUMA employs two gene-mapping strategies. First, it uses Multi-marker Analysis of Genomic Annotation (MAGMA) to aggregate SNP-level statistics up to the gene level, which enables more facile follow-up network analyses. However, MAGMA does not take gene regulatory information into account when mapping SNPs to genes. Alternatively, FUMA allows GWAS annotation by leveraging Hi-C and eQTL data, leveraging available data resources including GTEx, Brain eQTL Almanac (BRAINEAC) [93], CommonMind [94], and PsychENCODE.[95]

Hi-C-associated Multi-marker Analysis of GenoMic Annotation (H-MAGMA) was developed to overcome limitations in MAGMA.[96] H-MAGMA advances MAGMA by incorporating long-range (gene regulatory) interactions defined by Hi-C in mapping SNPs to genes. Further, it adopts the genome-wide mapping capability of MAGMA and expands the gene set to follow-up for molecular and biological pathway analysis. H-MAGMA has been developed on multiple Hi-C datasets[96,97]—those obtained from human fetal brains, adult brains, neurons, and glia sorted from the adult dorsolateral prefrontal cortex (DLPFC), iPSC-derived neurons, and iPSC-derived astrocytes. This enables developmental stage and cell type-specific gene mapping.

GeneWeaver is a suite of database and analysis tools that integrate data from expression microarray, RNA-seq, QTL mapping, GWAS, and mutation and perturbation screening experiments across species (yeast, worm, fly, zebrafish, mouse, rat, dog, human, and other species).[98-100] It also integrates protein-protein, molecular networks, and regulatory relationships to impute biological functions of variants and genes to phenotypes. In addition, GeneWeaver can assess molecular and trait relations through graphical network algorithms that leverage gene-gene and variant-variant comparison using complex, heterogeneous networks and random walk or network flow-based approaches. Until recently, GeneWeaver has used a gene-based strategy to integrate data because convergence or conservation of mechanism across species has typically relied on gene orthology. Authoritative data resources, including model organism databases and the Alliance of Genome Resources, have cataloged orthologous genes across species based on sequence alignments. Functional genomics analysis systems, including GeneWeaver, have made use of these reported orthologues to compare the results of genomic experiments across species at the gene level. Transitive associations are made to infer cross-species orthology where sequence alignment has not inferred a relationship (e.g., a *Drosophila*:zebrafish orthologue and zebrafish:mouse orthologue can be used to infer *Drosophila*:mouse orthology). Although functional coding variants, such as missense variants, are enriched among GWAS findings, most genome-wide significant variants implicate noncoding regions.[33] These noncoding variants are poorly conserved at the sequence level, and their functional interpretation presents a major challenge for the field. New approaches are being developed by the GeneWeaver project for mapping noncoding variants across species based on functional similarity and target orthology using combined genomic data sources. These methods are being applied to prioritize GWAS-identified variants based on evidence obtained in model organisms.

GeneNetwork. GeneNetwork is an interactive system for genome-to-phenome analysis, QTL mapping, and network integration. This resource incorporates large genetic, multi-omic, and phenotype data sets for highly diverse animal model populations such as the BXD and CC lines of mice, the HXB and HS rats, and several large number transcriptome data sets, including GTEx. GeneNetwork integrates 40 years of animal model data relevant to NIDA, NIAAA, NINDS, and NIMH missions, starting with catalytic studies by Crabbe, McClearn, Hitzemann

and Flint—especially data on behavioral variation and its linkage to gene and protein expression in the central nervous system.[55,68,101] The great majority of data in GeneNetwork is both open and FAIR-compliant and can be downloaded or used on-site in combination with powerful mapping modules that include R/qtl,[102,103] and the Bayesian Network Webserver.[104]

PrediXcan / MetaXcan. PrediXcan was developed as a gene-based association test that prioritizes genes likely to be causal for the phenotype, using predicted gene expression levels, most often with GTEx as the reference.[105] S-PrediXcan is a variation of this test that uses summary statistics instead of individual-level data. MultiXcan and S-MultiXcan are multivariate approaches (in contrast to the single-tissue approaches of PrediXcan/S-PrediXcan) that integrate measurements across tissues while accounting for correlations. Extensions of this approach are now being used to transfer polygenic findings from GWAS between human populations, and the authors suggest that these techniques might allow translation between species in the future.[106] These methods fall under the family of transcriptome-wide association study (TWAS)[107] approaches more broadly (e.g., Fusion is a similar approach that can be performed on GWAS summary statistics).[107]

Theme C: Ensuring that data are ready for integration

The long-term data curation and implementation of FAIR data principles (https://www.go-fair.org/fair-principles/) is integral to the success of integrating human and model organism research and multi-omic data. FAIR standards are particularly important. Without attention to data accessibility, many large and small SUD-related data sets risk evaporating over a relatively short period of time—often only five to ten years. This is particularly true of animal model data that tends to be highly granular and often siloed. Data sharing issues aside, there is a need for (inter)national storage and curation efforts because those aspects are typically beyond the scope of most research projects. Continued access to data, regardless of its presumed value, is key to leveraging future technological advances. There are, however, notable cases where advances in computing capacity and statistical methods greatly improve the value of older data. For example, phenotype data on drugs of abuse acquired over three decades ago can now be reanalyzed using new mapping algorithms (e.g., linear mixed models) and full genome sequence data. For example, data generated by a team at ORNL a

decade ago[68] can be remapped today to generate significantly stronger and even novel results than they did initially.

Participants discussed current knowledge gaps related to the development of metadata standards and data ontologies in order to move research forward. For instance, the lack of standards for describing disease phenotypes, such as those developed by the MONARCH initiative (Mondo disease ontology and Human Phenotype Ontology [HPO];[108,109]) and the limited amount and quality of derived phenotypes from electronic health records. Metadata helps with findability, interoperability, and usability. Because of this, participants emphasized that distribution platforms and curation tools that make metadata searchable urgently need further development. Overcoming these limitations would involve the identification of missing summary metadata fields for human data in dbGaP, as well as making prior results and data accessible both in name and in practice. Still, there is not a standard process for making data more findable and readable. Participants discussed several possible approaches for making data more searchable, such as using a Digital Object Identifier (DOI), machine-readable identification number, and Research Resource Identifiers (RRIDs)[110] as possible strategies to achieving data integration. As with all large-scale data endeavors, the researchers recognized a limitation around encryption software that would enable accessibility of primary raw data and allow searches across databases without the loss of de-identification. A major benefit of overcoming this limitation would be the ability to work with raw data using alternative methods that meta-analysis does not permit. Similarly, researchers acknowledge the limited number of Application Programming Interfaces (APIs) to enable interactions between data, applications, and devices. APIs deliver data and facilitate connectivity between devices and programs. Compelling prototype solutions are described above, but issues remain in the widespread integration and adoption of these systems. The biggest challenges are dynamic updating and organization of data for sharing and discovery as well as connecting across organisms and data types (e.g., sequence, epigenomic, etc.). Integration between graphical and relational databases remains a problem to be solved. To address these major challenges, participants discussed areas for improvement, including a lack of understanding of the following:

1. The degree of modularity and interoperability of existing data analysis software that can be used to facilitate the integration of ChIP-seq, DNA methylation, Hi-C, RNA-seq, splice variants, and structural variants information.
2. How gene network, epistasis, and genetic modifiers affect substance use outcomes.
3. How chromatin organization varies across human brain regions and in different cell types.
4. Ancestry differences in gene regulation.
5. How chromatin (Hi-C) and methylation (H3K27ac) data can be combined to predict gene expression with higher accuracy.
6. How models using protein-protein interaction (or similarly relevant omic data) data can help to improve the performance of existing genetic prediction tools.
7. How to access raw primary data while maintaining de-identification.

Conclusions and Future Directions

Genetics in human and animal models is now providing significant insights into molecular causes of addiction and SUDs. However, these leads still require extensive evaluation before being employed as prevention (e.g., to understand the utility of a polygenic score (PGS) beyond indicators of family history) and intervention tools (e.g., to reset CNS metabolic and cellular states back to health and well adapted behavior).[111] Major gaps in the field's mechanistic understanding of the perturbations underlying SUDs remain. Addressing these gaps and advancing the field will require attention to the following areas: (1) well-powered GWAS of SUDs and relevant human traits in diverse samples, (2) computational workflows that jointly leverage model organisms and large human cohorts, (3) generation and integration of multi-omic data across developmental stages, brain regions, molecularly defined cell types, and disease conditions, (4) data harmonization across human and model organisms at the level of the phenotype, as well as different omic, cellular, and systems levels, and (5) data curation and sharing.

Meeting participants also discussed key areas for future data integration, beginning with cross-species research and data integration tools. Continued research in integrative platforms will allow the examination of various use cases that will help develop an understanding of the

difficulties and opportunities in data integration. As the goal is to develop a plausible set of gene networks/sets from robust GWAS and fine mapping studies in mice and humans, it will be important to consider the nuances of mapping top results based solely on positional data. For example, previous SUD GWASs limited annotations to genes nearest to the lead SNP, and only more recently have studies begun to include tissue-specific annotation methods such as H-MAGMA and PrediXscan, to name a few. Many researchers are working on systematic multi-omic integration approaches to fine map complex genetic loci and nominate target genes. Reports on the progress of these efforts began at the Genetics and Epigenetics of Addiction (January 13–14, 2020) and are available at https://www.drugabuse.gov/research/research-data-measures-resources/genetics-epigenetics-ccrt/nida-genetics-consortium-ngc/nida-genetic-consortium-meetings-abstracts. Second, we need an increased understanding of the neurotoxic and behavioral effects of drugs. This continuously evolving body of literature will facilitate computational experiments to identify gene variants in underpowered GWAS. Integrative analyses in humans that include model organism data could also be applied to GWAS data as have been realized to date using Bayesian approaches to optimize gene identification using functional categories in genetics[112] and *cis*- and *trans*-eQTL information in transcriptomics.[113]

This Data Jamboree meeting represents a pivotal point in an ongoing process of information sharing that reflects the interdisciplinary nature of addiction genetics research. Notably, it builds on the previous report by Cates et al.,[114] that emphasized the importance of harmonizing phenotypes and methods of analysis among studies.

Even though geneticists at this meeting did not always agree on the ideal course of action for the next phase of discovery, the debate and dialogue, spurred by a shared commitment towards identifying tangible genetic targets, resulted in several new directions for human and model organism research.

Funding & Disclosure

The authors confirm that we have no conflicts of interest to declare. This work was supported by grants from the National Institute on Drug Abuse (DP1 DA042103 [awarded to: RHCP], P30DA044223 [LS & RW], R01 DA051913 [DBH & DJ]), R01 DA051908 [EOJ & DJ], U01 DA048279 [SA], R21 DA051921 [HW], DP1 DA044371 [JE], U01 DA041602 [DJS], P50DA037844 [AAP], DA028420 [MB], DA045401 [MB], K02 DA032573 [AA], R21 DA047527 [RP], R15 DA041618 [CCP], U24 DA039832 [MM], PGC-SUD support (MH109532 [ECJ, RP, JG, HJE, AA], and The University of Tennessee Center for Integrative and Translational Genomics [RW].

Acknowledgements

We would like to thank Drs. Amy Lossie, Jonathan Pollock, Susan Wright, Roger Little and Marti Head for their excellent organization of the workshop and their encouragement in assembling this report. We gratefully acknowledge Ms. Michelle Myers of RTI International for editorial assistance and Ms. Maia Amellio of Emory University for editorial assistance, as well as Dr. Megan Mulligan of The University of Tennessee Health Science Center for her assistance with the planning of the meeting.

Acknowledgements

The authors have no conflicts to declare.

Availability of Data

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

1. Our World in Data. Our world in data homepage. n.d.; https://ourworldindata.org/. Accessed October 29, 2020.
2. Centers for Disease Control and Prevention. Chronic liver disease and cirrhosis. 2021; https://www.cdc.gov/nchs/fastats/liver-disease.htm. Accessed February 11, 2021.
3. Centers for Disease Control and Prevention. Diabetes. 2021; https://www.cdc.gov/nchs/fastats/diabetes.htm Accessed October 29, 2020.
4. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet.* 2013;382(9904):1575-1586.
5. Goldman D, Oroszi G, Ducci F. The genetics of addictions: uncovering the genes. *Nat Rev Genet.* 2005;6(7):521-532.
6. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet.* 2013;14(6):379-389.
7. Lee CH, Eskin E, Han B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics.* 2017;33(14):i379-i388.
8. Zhu Z, Anttila V, Smoller JW, Lee PH. Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies. *PloS one.* 2018;13(3):e0193256.
9. Deak JD, Miller AP, Gizer IR. Genetics of alcohol use disorder: a review. *Curr Opin Psychol.* 2019;27:56-61.
10. Erzurumluoglu AM, Liu M, Jackson VE, et al. Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol Psychiatry.* 2020;25(10):2392-2409.
11. Jensen KP. A review of genome-wide association studies of stimulant and opioid use disorders. *Mol Neuropsychiatry.* 2016;2(1):37-45.
12. Johnson EC, Demontis D, Thorgeirsson TE, et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *The Lancet Psychiatry.* 2020;7(12):1032-1045.
13. Hancock DB, Markunas CA, Bierut LJ, Johnson EO. Human genetics of addiction: new insights and future directions. *Curr Psychiatry Rep.* 2018;20(2):8.
14. Zhou H, Sealock JM, Sanchez-Roige S, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat Neurosci.* 2020;23(7):809-818.
15. Walters RK, Polimanti R, Johnson EC, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat Neurosci.* 2018;21(12):1656-1669.
16. Sanchez-Roige S, Fontanillas P, Elson SL, et al. Genome-wide association study of alcohol use disorder identification test (AUDIT) scores in 20 328 research participants of European ancestry. *Addict Biol.* 2019;24(1):121-131.

17. Kranzler HR, Zhou H, Kember RL, et al. Author Correction: Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat Commun.* 2019;10(1):2275.
18. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet.* 2019;51(2):237-244.
19. Sanchez-Roige S, Palmer AA, Fontanillas P, et al. Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *Am J Psychiatry.* 2019;176(2):107-118.
20. Quach BC, Bray MJ, Gaddis NC, et al. Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat Commun.* 2020;11(1):5562.
21. Wendt FR, Pathak GA, Overstreet C, et al. Characterizing the effect of background selection on the polygenicity of brain-related traits. *Genomics.* 2021;113(1 Pt 1):111-119.
22. Zhou H, Rentsch CT, Cheng Z, et al. Association of OPRM1 functional coding variant with opioid use disorder: a genome-wide association study. *JAMA Psychiatry.* 2020.
23. Sun J, Kranzler HR, Gelernter J, Bi J. A genome-wide association study of cocaine use disorder accounting for phenotypic heterogeneity and gene-environment interaction. *J Psychiatry Neurosci.* 2020;45(1):34-44.
24. Sullivan PF, Agrawal A, Bulik CM, et al. Psychiatric genomics: an update and an agenda. *Am J Psychiatry.* 2018;175(1):15-27.
25. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421-427.
26. Wendt FR, Pathak GA, Overstreet C, et al. Natural selection influenced the genetic architecture of brain structure, behavioral and neuropsychiatric traits. *Biorxiv.* 2020.
27. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci.* 2019;22(3):343-352.
28. Converge Consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature.* 2015;523(7562):588-591.
29. Cai N, Revez JA, Adams MJ, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet.* 2020;52(4):437-447.
30. Deroche-Gamonet V. The relevance of animal models of addiction. *Addiction.* 2020;115(1):16-17.
31. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* 2015;16(4):197-212.
32. Gusev A, Lee SH, Trynka G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American journal of human genetics.* 2014;95(5):535-552.
33. Markunas CA, Johnson EO, Hancock DB. Comprehensive evaluation of disease- and trait-specific enrichment for eight functional elements among GWAS-identified variants. *Human genetics.* 2017;136(7):911-919.
34. Buchwald J, Chenoweth MJ, Palviainen T, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. *Mol Psychiatry.* 2020.

35. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav.* 2019;3(5):513-525.
36. Pritikin JN, Neale MC, Prom-Wormley EC, Clark SL, Verhulst B. GW-SEM 2.0: Efficient, Flexible, and Accessible Multivariate GWAS. *Behav Genet.* 2021.
37. Palmer RH, McGeary JE, Francazio S, et al. The genetics of alcohol dependence: advancing towards systems-based approaches. *Drug Alcohol Depend.* 2012;125(3):179-191.
38. Wainschtein P, Jain DP, Yengo L, et al. Recovery of trait heritability from whole genome sequence data. *bioRxiv.* 2019.
39. Wessel J, Majarian TD, Highland HM, et al. Rare non-coding variation identified by large scale whole genome sequencing reveals unexplained heritability of type 2 diabetes. 2020.
40. Brazel DM, Jiang Y, Hughey JM, et al. Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. *Biol Psychiatry.* 2019;85(11):946-955.
41. Kranzler HR, Zhou H, Kember RL, et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat Commun.* 2019;10(1):1499.
42. Byrne EM, Zhu Z, Qi T, et al. Conditional GWAS analysis to identify disorder-specific SNPs for psychiatric disorders. *Mol Psychiatry.* 2020.
43. Hivert V, Sidorenko J, Rohart F, et al. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *American journal of human genetics.* 2021.
44. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* 2019;51(4):584-591.
45. Bergen SE, Ploner A, Howrigan D, et al. Joint Contributions of Rare Copy Number Variants and Common SNPs to Risk for Schizophrenia. *Am J Psychiatry.* 2019;176(1):29-35.
46. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell.* 2019;179(3):589-603.
47. Skene NG, Bryois J, Bakken TE, et al. Genetic identification of brain cell types underlying schizophrenia. *Nat Genet.* 2018;50(6):825-833.
48. Hatoum AS, Wendt FR, Galimberti M, et al. Genetic data can lead to medical discrimination: opioid use disorder as a cautionary tale. 2020.
49. Shi X, Walter NA, Harkness JH, et al. Genetic polymorphisms affect mouse and human trace amine-associated receptor 1 function. *PLoS One.* 2016;11(3):e0152581.
50. Juraeva D, Treutlein J, Scholz H, et al. XRCC5 as a risk gene for alcohol dependence: evidence from a genome-wide gene-set-based analysis and follow-up studies in Drosophila and humans. *Neuropsychopharmacology.* 2015;40(2):361-371.
51. Homanics GE. Gene-edited CRISPy Critters for alcohol research. *Alcohol.* 2019;74:11-19.
52. Mackay TF, Richards S, Stone EA, et al. The Drosophila melanogaster genetic reference panel. *Nature.* 2012;482(7384):173-178.

53. Huang W, Massouras A, Inoue Y, et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res.* 2014;24(7):1193-1208.
54. Williams RW, Williams EG. Resources for systems genetics. *Methods Mol Biol.* 2017;1488:3-29.
55. Ashbrook DG, Arends D, Prins P, et al. A platform for experimental precision medicine: the extended BXD mouse family. *Cell Syst.* 2021.
56. Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. *Mamm Genome.* 2012;23(9-10):713-718.
57. Hansen C, Spuhler K. Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcohol Clin Exp Res.* 1984;8(5):477-479.
58. Mott R, Flint J. Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. *Genetics.* 2002;160(4):1609-1618.
59. Valdar W, Solberg LC, Gauguier D, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet.* 2006;38(8):879-887.
60. Solberg Woods LC, Palmer AA. Using heterogeneous stocks for fine-mapping genetically complex traits. *Methods Mol Biol.* 2019;2018:233-247.
61. Gileta AF, Fitzpatrick CJ, Chitre AS, et al. Genetic characterization of outbred Sprague Dawley rats and utility for genome-wide association studies. *bioRxiv.* 2018.
62. Fitzpatrick CJ, Gopalakrishnan S, Cogan ES, et al. Variation in the form of Pavlovian conditioned approach behavior among outbred male Sprague-Dawley rats from different vendors and colonies: sign-tracking vs. goal-tracking. *PLoS One.* 2013;8(10):e75042.
63. Mulligan MK, Abreo T, Neuner SM, et al. Identification of a functional non-coding variant in the GABA (A) receptor α2 subunit of the C57BL/6J mouse reference genome: major implications for neuroscience research. *Front Genet.* 2019;10:188.
64. Stafford AM, Reed C, Baba H, et al. Taar1 gene variants have a causal role in methamphetamine intake and response and interact with Oprm1. *Elife.* 2019;8.
65. Dodd S, A FC, Puri BK, et al. Trace Amine-Associated Receptor 1 (TAAR1): A new drug target for psychiatry? *Neurosci Biobehav Rev.* 2021;120:537-541.
66. Jones P, Weighill D, Shah M, et al. Network modeling of complex data sets. *Methods Mol Biol.* 2020;2096:197-215.
67. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
68. Philip VM, Duvvuru S, Gomero B, et al. High-throughput behavioral phenotyping in the expanded panel of BXD recombinant inbred strains. *Genes Brain Behav.* 2010;9(2):129-159.
69. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell.* 2019;177(4):1080.
70. Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. *Science.* 2016;354(6308):54-59.
71. Sittig LJ, Carbonetto P, Engel KA, Krauss KS, Barrios-Camacho CM, Palmer AA. Genetic Background Limits Generalizability of Genotype-Phenotype Relationships. *Neuron.* 2016;91(6):1253-1259.
72. Palmer RHC, Benca-Bachman CE, Huggett SB, et al. Multi-omic and multi-species meta-analyses of nicotine consumption. *Transl Psychiatry.* 2021;11(1):98.

73. Farris SP, Riley BP, Williams RW, et al. Cross-species molecular dissection across alcohol behavioral domains. *Alcohol.* 2018;72:19-31.
74. Mignogna KM, Bacanu SA, Riley BP, Wolen AR, Miles MF. Cross-species alcohol dependence-associated gene networks: Co-analysis of mouse brain gene expression and human genome-wide association data. *PloS one.* 2019;14(4):e0202063.
75. The Genotype-Tissue Expression (GTEx) Project https://gtexportal.org/home/. Accessed 03/28/2021.
76. van Berkum NL, Lieberman-Aiden E, Williams L, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp.* 2010(39).
77. Cliff A, Romero J, Kainer D, Walker A, Furches A, Jacobson D. A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes (Basel).* 2019;10(12).
78. Hatoum AS, Wendt FR, Galimberti M, et al. Genetic Data Can Lead to Medical Discrimination: Cautionary tale of Opioid Use Disorder. *medRxiv.* 2020:2020.2009.2012.20193342.
79. Furches A, Kainer D, Weighill D, et al. Finding new cell wall regulatory genes in populus trichocarpa using multiple lines of evidence. *Front Plant Sci.* 2019;10:1249.
80. Joubert W, Weighill D, Kainer D, et al. Gordan Bell Prize Winner: attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction. SC18: International Conference for High Performance Computing, Networking, Storage and Analysis; November, 2018.
81. Valdeolivas A, Tichit L, Navarro C, et al. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics.* 2019;35(3):497-505.
82. Lane SP, Sher KJ. Limits of current approaches to diagnosis severity based on criterion counts: an example with DSM-5 alcohol use disorder. *Clin Psychol Sci.* 2015;3(6):819-835.
83. Palmer RHC, Brick LA, Chou YL, et al. The etiology of DSM-5 alcohol use disorder: Evidence of shared and non-shared additive genetic effects. *Drug and alcohol dependence.* 2019;201:147-154.
84. Palmer RH, McGeary JE, Heath AC, Keller MC, Brick LA, Knopik VS. Shared additive genetic influences on DSM-IV criteria for alcohol dependence in subjects of European ancestry. *Addiction.* 2015;110(12):1922-1931.
85. Lai D, Wetherill L, Bertelsen S, et al. Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes Brain Behav.* 2019;18(6):e12579.
86. Marees AT, Smit DJA, Ong JS, et al. Potential influence of socioeconomic status on genetic correlations between alcohol consumption measures and mental health. *Psychol Med.* 2020;50(3):484-498.
87. Johnson EC, Demontis D, Thorgeirsson TE, et al. A large-scale genome-wide association study meta-analysis of cannabis use disorder. *The lancet Psychiatry.* 2020;7(12):1032-1045.
88. Palmer RH, Benca-Bachman C, Huggett S, et al. Cross-species integration of transcriptomic effects of tobacco and nicotine exposure helps to prioritize genetic effects on human tobacco consumption. *Translational Psychiatry.* in press.

89. van Swinderen B, Greenspan RJ. Flexibility in a gene network affecting a simple behavior in Drosophila melanogaster. *Genetics.* 2005;169(4):2151-2163.
90. Huggett SB, Stallings MC. Cocaine'omics: Genome-wide and transcriptome-wide analyses provide biological insight into cocaine use and dependence. *Addiction biology.* 2020;25(2):e12719.
91. Reynolds T, Johnson EC, Huggett SB, et al. Interpretation of psychiatric genome-wide association studies with multispecies heterogeneous functional genomic data integration. *Neuropsychopharmacology.* 2021;46(1):86-97.
92. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.
93. BRAINEAC: Web server for data from the UK Brain Expression Consortium (UKBEC). http://www.braineac.org/. Accessed 03/28/2021.
94. Hoffman GE, Bendl J, Voloudakis G, et al. CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data.* 2019;6(1):180.
95. Psych EC, Akbarian S, Liu C, et al. The PsychENCODE project. *Nat Neurosci.* 2015;18(12):1707-1712.
96. Sey NYA, Hu B, Mah W, et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat Neurosci.* 2020;23(4):583-593.
97. Hu B, Won H, Mah W, et al. Neuronal and glial 3D chromatin architecture illustrates cellular etiology of brain disorders. 2020.
98. Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ. GeneWeaver: a web-based system for integrative functional genomics. *Nucleic acids research.* 2012;40(Database issue):D1067-1076.
99. Bubier JA, Langston MA, Baker EJ, Chesler EJ. Integrative Functional Genomics for Systems Genetics in GeneWeaver.org. *Methods Mol Biol.* 2017;1488:131-152.
100. Bubier J, Hill D, Mukherjee G, et al. Curating gene sets: challenges and opportunities for integrative analysis. *Database (Oxford).* 2019;2019.
101. Chesler EJ, Lu L, Wang J, Williams RW, Manly KF. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat Neurosci.* 2004;7(5):485-486.
102. Finazzo MS, Hoffman MS, Roberts WS, Cavanagh DM. Previous pelvic surgery in patients with ovarian cancer. *South Med J.* 1988;81(12):1518-1520.
103. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44(7):821-824.
104. Ziebarth JD, Cui Y. Precise network modeling of systems genetics data using the Bayesian network webserver. *Methods Mol Biol.* 2017;1488:319-335.
105. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 2019;15(1):e1007889.
106. Liang Y, Pividori M, Manichaikul A, et al. Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries. *Biorxiv.* 2020. https://www.biorxiv.org/content/10.1101/2020.11.12.373647v1?rss=1.

107. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* 2016;48(3):245-252.
108. Yuan S, Stratton CJ, Bao J, et al. Spata6 is required for normal assembly of the sperm connecting piece and tight head-tail conjunction. *Proc Natl Acad Sci U S A.* 2015;112(5):E430-439.
109. Mungall CJ, McMurry JA, Kohler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* 2017;45(D1):D712-D722.
110. Bandrowski AE, Martone ME. RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. *Neuron.* 2016;90(3):434-436.
111. Fullerton JM, Nurnberger JI. Polygenic risk scores in psychiatry: Will they be useful for clinicians? *F1000Res.* 2019;8.
112. Yang J, Fritsche LG, Zhou X, Abecasis G. International age-related macular degeneration genomics C. A scalable Bayesian method for integrating functional information in genome-wide association studies. *American journal of human genetics.* 2017;101(3):404-416.
113. Luningham JM, Chen J, Tang S, et al. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *American journal of human genetics.* 2020;107(4):714-726.
114. Cates HM, Benca-Bachman CE, de Guglielmo G, Schoenrock SA, Shu C, Kallupi M. National Institute on Drug Abuse genomics consortium white paper: Coordinating efforts between human and animal addiction studies. *Genes Brain Behav.* 2019;18(6):e12577.
115. Consortium PG. Psychiatric Genomics Consortium. https://www.med.unc.edu/pgc/. Accessed 03/28/2021.
116. Genetics d. deCODE Genetics. https://www.decode.com/. Accessed 03/28/2021.
117. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12(3):e1001779.
118. Hivert V, Sidorenko J, Rohart F, et al. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *bioRxiv.* 2020.

932

Table 1. Considerations and Areas of Opportunity for Data Integration				
Methodological Approach	**Considerations in Model Organism Genetics**	**Considerations for Human Genetics**	**Considerations for Reductionist Models (Human and Model Organisms)**	**Areas of Convergence**
G x E	Many populations provide favorable recombination and allele frequencies to provide adequate power to detect G x E effects	Consortia efforts (e.g., Psychiatric Genetics Consortium (PGC) [115], deCODE Genetics [116], UK biobank [117], etc.) and integration of electronic health records can help construct large sample sizes for improved power to detect G x E effects	Not possible to mimic most environmental effects (e.g. social interactions, early life adversity, etc) in cell lines or organ cultures	-Animal models can test the effects of a specific gene implicated in human GWAS across multiple environments, or different genes in the same environment. -G x E hits from QTL mapping can be used to prioritize promising variants in human GWAS that did not meet significance thresholds due stringent corrections for multiple testing
	Some human environments are not possible to	Some environments are unethical to		

61
62

	model in animals	impose on humans		
G x G	QTL mapping in many populations can provide sufficient power to examine other forms of DNA variation and potential nonlinear G x G effects	Need very large sample sizes (> 1 million) to detect potential nonlinear G x G effects[118]	QTL mapping efforts should utilize genetically diverse populations in order to better extrapolate results across strains and species	-Development of new statistical models to detect G x G epistatic interactions will improve our understanding of the polygenic nature of SUDs -Use of genetically admixed, mutant, and genetically simple cohorts of model organisms can identify epigenetic modifiers
	Structured panels of F_1 progeny that place null alleles on different genetic backgrounds can identify G x background interactions	Consortia efforts and private Direct to Consumer biotechnology companies (e.g. 23 & me , ancestry.com) may be key to amassing large enough sample sizes for improved power to detect epistasis	If using CRISPR to study G x G interactions, researchers should test multiple genetic backgrounds	
	CRISPR allows for simultaneous			

	alteration of multiple genes to examine G X G interactions			
Meta-analysis	Not commonly performed in model organisms, but the extendable nature of many populations is favorable to this approach	Meta-analysis has been key in the successful identification and replication of loci across human studies, thus increasing power and reproducibility		-Development and application of metadata standards and data ontologies (such as MONARCH) will be critical to harmonize data across organisms and data types -Improved data curation and sharing will allow for increased accessibility to all researchers -Meta-analytic studies using omics data from both mapping populations and mutant animals can detect and validate novel findings entirely *in silico*
Polygenic Risk Scores	Must account for allele frequency differences across	Must account for allele frequency differences across		-Need to develop methodology to integrate PGS between animals and

	populations	populations		humans to improve translational, predictive and clinical utility
	Not widely implemented in animal QTL mapping studies	PGS in humans have allowed cross-trait and cross-sample comparisons, greatly enhancing our knowledge of SUDs		
	For translational studies, need to limit PGS variants to those with orthologs in humans			
Proteomics/ Transcriptomics	Can be easily obtained in animals from relevant tissues, cell-types, and timepoints (post-drug, developmental)	Post-mortem brain tissue from humans is confounded by life histories, drug use patterns, time elapsed between death and brain collection		-Multi-omics data (genome, epigenome, transcriptome, proteome, metabolome, microbiome) data in both model organisms and humans can improve our understanding of GWAS hits that fall in regulatory regions -Single-cell RNAseq will help

	Multiple bioinformatics resources exist to integrate omics results (GeneWeaver, GeneNetwork)	Web-based repositories (GTEx, BRAINEAC, CommonMind, PsychENCODE) provide valuable resources to examine effects of gene expression on disease		uncover cell-type specific networks involved in SUDs -Animal models may identify mobile element polymorphisms, inversions, and other structural variants that can later be studied in human GWAS -Network integration (such as LOE, RWR) is key to permit the full illumination of patterns shared across multi-omics datasets and can be used to leverage information across species -Exploiting publicly available bioinformatics resources can provide secondary study replication/validation, increase power, and provide *a priori* information for study hypotheses and design
Functional Validatio	Multiple genetic resources exist (CRISPR, KO,	Unethical to perform gene editing studies in	Functional validation studies should test the	-Model organisms provide opportunities to test the effects of a specific gene(s)

n	transgenics, RNAi, etc) to functionally validate genes of interest in developmental-, tissue-, and cell-specific regions	humans	effects of gene manipulation on multiple genetic backgrounds	implicated in human GWAS to help elucidate the underlying biology -Functional validation studies may benefit from cross-species analysis (yeast, worms, flies allow for the analysis of hundreds of candidate genes) -Development of efficient and unbiased computational workflows (such as FUMA GWAS, H-MAGMA, GeneWeaver, PrediXcan/MetXcan) is needed to rank top variants and map their cellular networks in both human and model organisms
	Optogenetic and other brain stimulation approaches can isolate neurons, define pathways relevant to traits of interest	Transcranial magnetic stimulation can excite/silence brain regions in humans, but is limited		
	Lesion studies can readily be performed in animal models	Naturally occurring lesions can be studied		
Environmental Control	Can more tightly control environmental parameters	Diverse environmental and lifestyle influences		-Improved statistical models that better account for confounds, Winner's Curse, and cofactors/covariates will enhance translational potential for both animal and
	Cannot accurately	Differing		

71
72

	model some human components (e.g., social elements) of environments	combinations of psychiatric and other risk factors		human research