Interpretation of psychiatric genome wide association studies with multi-species heterogeneous functional genomic data integration

Cite this article as: Timothy Reynolds, Emma Johnson, Spencer Huggett, Jason A. Bubier, Rohan H. C. Palmer, Arpana Agrawal, Erich J. Baker and Elissa J. Chesler, Interpretation of psychiatric genome wide association studies with multi-species heterogeneous functional genomic data integration, *Neuropsychopharmacology* doi:10.1038/s41386-020-00795-5

Neuropsychopharmacology Review

Interpretation of psychiatric genome wide association studies with multi-species heterogeneous functional genomic data integration

Timothy Reynolds[1,2], Emma Johnson[3], Spencer Huggett[4], Jason A. Bubier[1], Rohan H.C. Palmer[4], Arpana Agrawal[3], Erich J. Baker[2], *Elissa J. Chesler[1]

[1]The Jackson Laboratory
[2]Computer Science Department, Baylor University
[3]Department of Psychiatry, Washington University St Louis
[4]Emory University

*Corresponding Author

The Jackson Laboratory
600 Main Street
Bar Harbor ME 04609
207-288-6453
Elissa.Chesler@jax.org

Abstract

Genome-wide association studies and other discovery genetics methods provide a means to identify previously unknown biological mechanisms underlying behavioral disorders that may point to new therapeutic avenues, augment diagnostic tools and yield a deeper understanding of the biology of psychiatric conditions. Recent advances in psychiatric genetics have been made possible through large-scale collaborative efforts. These studies have begun to unearth many novel genetic variants associated with psychiatric disorders and behavioral traits in human populations. Significant challenges remain in characterizing the resulting disease-associated genetic variants and prioritizing functional follow-up to make them useful for mechanistic understanding and development of therapeutics. Model organism research has generated extensive genomic data that can provide insight into the neurobiological mechanisms of variant action, but a cohesive effort must be made to establish which aspects of the biological modulation of behavioral traits are evolutionarily conserved across species. Scalable computing, new data integration strategies, and advanced analysis methods outlined in this review provide a framework to efficiently harness model organism data in support of clinically relevant psychiatric phenotypes.

PROMISES AND CHALLENGES IN HUMAN GENETICS OF PSYCHIATRIC DISORDERS

Psychiatric disorders are highly polygenic and show a continuous range of variation influenced by both environmental and genetic factors [1]. A major goal of psychiatric genetic research is to better understand the molecular mechanisms through which genetic variants act to influence liability to these traits. The identification of novel genetic variants provides a foothold into the complex genetic architecture that undergirds psychiatric traits. Model organisms provide an avenue into understanding the biological mechanisms that are impacted by genetic variation. In this review, we outline Big Data approaches that efficiently weave the vast amounts of convergent genomic data from other species into human genetic findings to elevate the likelihood of uncovering biologically meaningful pathways for further experimental follow-up and therapeutic discovery.

The utility of GWAS in psychiatry

Genome-wide association studies (GWAS) of psychiatric traits have generated an outpouring of recent discoveries in risk variant identification and polygenic prediction. From highly heritable traits, such as schizophrenia (for which >100 common loci have been reported with N = 150,064 [2]) to common but less heritable conditions such as problematic alcohol use (for which 29 independent loci have been reported with N = 435, 563 [3]) and major depression (for which 102 common loci were detected with N = 807,553 [4]), as well as for liability across psychiatric disorders (109 loci with N = 727,126 [5]) progress abounds. In addition, for substance use, a recent large GWAS of tobacco smoking (N for smoking initiation = 1,232,091) and typical drinking (N for drinks/week = 941,280) has identified over 400 loci [6].

The increased power accumulated across studies of major psychiatric disorders, arising from collaborative research, has revealed clues into novel mechanisms of susceptibility to mental illnesses and substance use disorders. These large-scale GWAS have also revealed patterns of genetic variation associated with multiple disorders as well as disorder-specific loci, e.g. *CADM2* has been linked to multiple substances and common addiction mechanisms (e.g., risk-taking cognition), while the alcohol dehydrogenase genes remain alcohol-specific (e.g. [7], [8]).

Challenges and opportunities within GWAS for psychiatric genetic studies

The recent gains in psychiatric genetic studies outlined above amplify the need to address several enduring challenges within GWAS. First, at a variant level, the bulk of GWAS "hits" fall in non-coding regions of the genome. A major advantage of GWAS as a means of discovering the biological basis of psychiatric disorders is that the lack of *a priori* gene centric hypotheses enables discovery of trait regulatory variants in enhancer and promotor regions, lncRNAs, microRNAs and any other molecular entity that is part of the gene regulatory mechanism. However, in contrast to variants within coding genes, it is far more difficult to link statistically significant genetic associations to the gene products and biological mechanisms through which they act [9]. Interpretations of significant GWAS findings are complicated by patterns of related inheritance (e.g., linkage disequilibrium), such that the most significant genetic variant in a locus may not be "causal" but could "tag" a true causal variant. This, coupled with long distance genomic regulation, poses challenges for unveiling specific genes and variants underlying human traits via GWAS [10]. In this review, we highlight how regulatory genetic variants can be integrated coherently with coding genes within and across species using unifying data structures.

A second challenge with GWAS is that power analyses reveal that the massive polygenicity underlying psychiatrically relevant traits and illnesses requires larger sample sizes for additional discoveries from GWAS data alone [11]. Likewise, the predictive power of a polygenic risk score (PRS), an index of aggregated genetic susceptibility to a disorder, for psychiatric disorders is also directly linked to the current statistical power of discovery GWAS [12]. However, the identification of additional trait-associated variants continues to substantially augment SNP-heritability estimates, especially in the case of rare variants, suggesting that there is more signal to be found in GWAS and sequencing studies [13], provided that higher sample sizes continue to be attained. In this review, we highlight approaches that exploit complementary data resources from model organisms that, when placed in an integrative framework with GWAS data, are showing some promise in prioritizing variants that are detected.

Third, consistent with indications from early family and twin studies, there is evidence for pleiotropy among psychiatric traits to a degree suggestive of an underlying dimension of genetic liability that parallels the general factor model of psychopathology [5], [14]. Thus, it is important to consider variants in context of both the underlying neurobiological mechanisms in which they function, and the multiple traits which are influenced by that variation to find the specific, as well as the overlapping biological mechanisms underlying behavioral traits. A landmark contribution to our current ability to annotate GWAS signals arise from FUMA [15], a platform for functional and regulatory annotation of variants. Summary statistics from a GWAS can easily be aligned with tissue and cell-type specific expression data and to a variety of regulatory and chromatin signatures with no computational burden on the user, making FUMA widely accessible. As an alternative to gene-based mapping techniques, software tools can also map variants to the non-coding transcriptome (e.g., LincSNP 3.0 [16]). Beyond variant mapping,

harnessing multiple sources of omics data can be utilized in a multivariate framework to implicate "causal" gene sets for a disease state (e.g., SMR [17], iRIGs [18], PAINTOR [19], FOCUS [20]). Efforts are also underway, with varying degrees of success, to demonstrate to what extent similar regulatory enrichment of polygenic risk scores could enhance prediction (e.g., AnnoPred [21], LDpred-funct [22]). However, most of these approaches have been limited to human genetics and genomics data. In this review, we highlight approaches that bring together the breadth and depth of well-controlled model organism studies that place genetic and genomic findings in biobehavioral context that can expand on this or other interpretive tool sets.

MULTI-SPECIES GENOMICS TO ADDRESS CHALLENGES IN GWAS VARIANT INTERPRETATION

Across these historical and contemporary research challenges, Big Data approaches that harness information from additional sources, including cross-species genomic analyses, can provide elegant solutions to current barriers in psychiatric genetics [18], [23]. It cannot be understated that we need better-powered GWAS, especially as we look to polygenic scores as a means of leveraging the modest effect sizes from GWAS. However, increasing the sample size alone may be merely a theoretical solution for certain traits where rare variation and modest effect sizes contribute substantially. Incorporating evidence from molecular and cellular biology shifts the focus of genome-wide analyses from variant detection and identification to evaluating the relative contribution of a prioritized subset of loci. This helps control the familywise error rate, thus increasing power, and provides context about the genome at multiple levels (i.e., structure, function, and regulation) while also accounting for the polygenicity of a trait.

Leveraging information from annotated genomic regions that affect gene function was shown to robustly increase the power to identify genomic associations across 27 human traits [24].

There is extensive information available from human and model organism functional genomics that may be brought to bear on human GWAS findings in the context of specific behaviors, tissues, and molecular mechanisms [25], [26]. Prior to the widespread availability of human ‘omics data, some of the earliest efforts to characterize the mechanism of variants detected in human association studies relied on expression of orthologous genes from studies performed in animal models. The rich data resources from these studies continue to be valuable due to the breadth and depth of studies that are possible in animal models, under precisely controlled conditions of drug exposure and other neurobiological or behavioral processes. Further, model organism data also contains a rich source of expression regulatory information including eQTL and epigenetic data from many tissues and brain regions, some of which is collected in populations that facilitate the global correlation of transcript abundance to neurobiological and behavioral parameters [27]. Integration of functional genomic information from multiple species into GWAS provides new clues about the biological context and consequences of genetic associations and polygenic risk scores, and provides insight into how to model such variation in *in vivo* preclinical models with intact central nervous systems and expression regulatory machinery.

Below, we illustrate the promise of harnessing these model organism data, for which decades of comparative behavioral research has produced numerous experimental paradigms aimed at consilience, such as drug self-administration and response studies across multiple mouse and rat populations in genetics and genomics [28]. We propose methods for integrating valuable and ever-expanding complementary model organism and human genetics and genomics

data (such as GTEx [29] and GeneNetwork.org [30], psychENCODE [31] and modENCODE [32]) and highlight new approaches for boosting power in human genetics through Bayesian inference in heritability and polygenic analyses, outline exciting developments aimed at bridging the "analytic currency" gap between human and model organism research, and present some technical and philosophical challenges. The overarching goal of this review is to focus on ways in which we might utilize the complementary strengths of human and animal genetics to advance their common research mission: gaining a better understanding of the biology of complex traits.

Potential and Challenges for Model Organism Data Integration

There is considerable and growing interest in employing non-human animal models to meet some of the challenges for human genetics outlined above. There is a tremendous depth and breadth of model organism genetics and genomics studies spanning many areas of behavioral and neurobiological parameters. These include differential expression studies following various behavioral and drug exposure paradigms [33], large-scale screens of gene-targeted deletion mutants [34], and genetic studies in populations such as the BXD RI mouse lines [35] and inbred strain panels [36] which often combine gene expression and genetic analysis. Numerous QTL positional candidates have been identified from a large number of behavioral and neurobiological mapping studies [37]. Selective breeding in rats and mice have been able to separate alcohol preferences [38], [39] and chronic use/withdrawal [40]. These data provide a rich backdrop and context in which to interpret the more global phenotype or disease information that is the frequent subject of GWAS analysis.

Animal geneticists have a rich history of using model organisms to study behavioral traits that mirror aspects of human psychopathology. Many of the genes and variants identified in model organisms are also now also being found in human GWAS studies (Table 1), indicating that convergence of these studies is feasible. To date, model organism evidence has largely been used as a form of post-GWAS validation to characterize significant SNP/gene effects (e.g., [41], [42]). There have been a few promising recent examples of model organism research that, when coupled with human GWAS findings, have revealed insights into the biological mechanisms underlying psychiatric disorders. Model organism data has also produced experimental insight into disease mechanism. For example, researchers used mouse models to study the effect of a particular protein, complement component 4 (C4), on synaptic mediation during development [43]. By using a mouse model in conjunction with convergent evidence from human genomic studies, researchers were able to study the effects of *C4* gene deficiencies on synapse elimination during post-natal development in a way that is not possible in humans. Researchers are beginning to leverage model organism genomics directly in the context of human genetic studies. For instance, gene co-expression networks associated with mouse neurodegeneration phenotypes demonstrated enrichment for human GWAS associations with Alzheimer's Disease [44]. Integrative methods for jointly analyzing model organism data directly with human GWAS are are under active development. One recent example identified novel brain mechanisms of alcohol use and dependence by co-analyzing human GWAS, human protein-protein interaction networks and mouse gene co-expression data. In doing so, the researchers interrogated ethanol-responsiveness genes obtained from mouse gene expression data of the PFC, VTA, and NAc [45].

Despite this substantial progress, there remain conceptual and technical challenges for data integration across species. These occur at the levels of phenotypic comparison, genetic conservation, and computational scale. A major challenge at the phenomic level is that any effort to integrate evidence across model organisms and humans must acknowledge that human psychiatric diagnoses and classifications are often based upon clinical instruments and nosology that are not easily transferable to model organisms, therefore efforts to "diagnose" animal models are discouraged. However, it is apparent that aspects of a disorder can transfer across species and be easily captured with experimental data, and increasingly, GWAS of psychiatric disorders are providing corroborating support for variants that influence both disorders and their trait-like manifestations that may be recapitulated in model organisms [46]. For example, it was recently shown that ethanol responsive genes in mouse prefrontal cortex, nucleus accumbens and ventral tegmental area were overrepresented in GWAS for alcohol dependence in the Irish Affected Sib-Pair Study of Alcohol Dependence and the Avon Longitudinal Study of Parents and Children [26]. The identification of network-level associations between humans and mice suggests shared sensitivity in ethanol responding, and thus can serve as support for nominal GWAS signals. However, far more complexity and heterogeneity than ethanol response underlies alcohol dependence in humans. Recent genomic distinctions identified between the consumption (AUDIT-C items 1-3) and the problematic (AUDIT-P items 7-10) subscales of the Alcohol Use Disorder Inventory Test (AUDIT) [8], [47] echo similar findings in model systems, the data from which will be critical for the interpretation of molecular mechanisms [48].

There is concern that comparative, multi-species approaches will not be as readily feasible for certain psychiatric traits. Behavioral characteristics including speech, language and certain executive and metacognitive functions are also impossible to assess in model organisms.

However, most studies that attempt comparative genomics across species are based on limited genetic diversity, often comparing a single idiosyncratic strain to a small sample of the population of humans, e.g. [49], and therefore can not discern between individual differences within populations and between species. For some disorders, there is a substantial role of brain structures that are under developmental control of poorly conserved genomic regions, leading to significant cross-species differences in these structures [50]. This potentially could preclude detection of genetic variants which regulate disorders through effects on the development of these structures. Following this logic, some aspects of substance use disorders are served by neural structures that show more conservation and may be more likely to provide convergent mechanistic evidence for overt characteristics of drug intake, withdrawal, compulsive responding even with choice and punishment, but perhaps not "desire to quit" or other metacognitive and psychosocial aspects of addiction.

However, all psychiatric disorders including SUDs are highly complex traits likely involving many risk loci. Some of these traits are manifest across species, even if the end-result in humans includes behavioral output not readily observable in non-human model organisms. Therefore, one can model the effects of genetic risk variants on more proximal biological consequences; for example, one might study the influence of C4 variation [43] on endophenotypes captured in Research Domain Criteria (RDoC) including synaptic excitability, or neuronal reactivity and the various startle phenotypes it is associated with, but not all of the species specific cognitive and behavioral output that are central to the disease pathology. Historically, the field has been distracted by pharmacologically predictive characteristics that have little face validity with the disorders to which they are applied [51]. Below we describe how cross-species comparative genomics provides a tool that can be used to identify what aspects of

the human disorder are reflected in model organism genomics, allowing data-driven discovery of the relations among traits across species [52].

At the genetic conservation level, cross-species genetic research has been hindered by the "analytic currency" problem. Human geneticists typically work at the variant level, and genomics data, particularly from expression studies, are often reported at the gene or transcript level. Prior efforts at model organism follow-up of human GWAS data were limited to human variants that could be positionally assigned to a gene, but as described below, this is no longer the case. As is evident from regulatory mapping analyses, the action of a variant does not readily correspond to the most proximal gene, or even a single gene. Further compounding the problem, non-coding regulatory variants are often found in poorly conserved regions of the genome, which renders cross-species gene orthology mapping challenging and variant mapping through sequence alone, impossible in many cases. Therefore, approaches that exploit both gene orthology and convergence of variant regulatory relations are most promising toward relating trait regulatory variation across species.

In the case of intragenic variants, current methods use transcript and protein annotations to identify causal SNPs based on the severity of mRNA and protein modifications [53] and other functional consequences [54]. However, the majority of SNPs are intergenic, suggesting the involvement of distal gene-regulatory mechanisms (e.g., chromatin accessibility). Therefore, the common approach of associating SNPs to nearby downstream and upstream genes can elicit false positives [55] and therefore it is necessary to use data from gene expression quantitative trait loci (eQTLs), epigenetics and 3D genomics to assess the relationships among regulatory variants and their distal targets.

Although most prior variant-to-gene annotation efforts have relied on positional approaches, i.e., assigning SNPs to genes based solely on physical proximity (e.g., MAGMA software [56]), modern approaches in humans rely on extensively curated functional and regulatory mapping from 'omics data (e.g., S-PrediXcan software [57], TWAS [58], Hi-C coupled MAGMA or H-MAGMA software [59]).

However, all of these approaches have almost exclusively used data from human genomic analyses. Similar approaches have been deployed in model organisms, but the integration of resources across species has remained rather incomplete, limiting the approach to a small number of applications. To facilitate cross-species analysis, integrative data analyses have historically relied on gene homology associations from model organism databases [60] and gene orthology services [61]. Analysis involving multiple species therefore most often occurs at the gene level, introducing a GWAS-specific integration challenge: the need to associate genetic variants with genes. For complex disorders, such as schizophrenia and SUDs, this often requires characterization of the regulatory nature of genetic variants associated with disease, or identifying functional variants in sub-molecular domains of drug targets that could confer vulnerability or resistance to various treatment. However, non-coding regions of the genome are often very poorly conserved across species, and the targets of the variants can be far away. Moreover, many of the implicated non-coding variants in GWAS reside in gene expression regulatory regions [62]. Here, we highlight solutions for the assessment of conserved effects of variants through their orthologous genomic targets to support a wide-range of applications in integrative functional genomics (Figure 1).

SOLUTIONS FOR DATA-DRIVEN CROSS-SPECIES ANALYSIS

Broadly speaking, integration of multi-species functional genomic data can occur in two ways—from the phenomic or genomic orientation. For example, top-down, trait-based approaches to cross-species analysis utilize the similarity of human disease-related phenotypic profiles to model organism phenotypic profiles to identify gene-disease associations [63]. These approaches, embodied in resources developed by The Monarch Initiative [64] identify similar phenotypes across species through integrated ontologies and semantic similarity methodologies that apply semantic reasoners to a unified knowledge graph [65]. Such phenotype-driven approaches, which leverage multi-species data, have been effective at assisting rare disease diagnosis [66] and improving identification of causal genetic mechanisms [67], but these approaches are challenging to apply in the context of high phenotypic and genetic heterogeneity due to the extensive differences among species in the behavioral manifestations of neurobiological variation.

In highly complex psychiatric disorders in which model organism traits may only capture a facet of the human disease, alternative bottom-up strategies that aggregate genomic data may be more suitable for identification of the driving genetic mechanisms associated with complex traits and disease. The varieties of biological entities—genes, proteins, variants, methylation sites, and chromatin states for example, which can be characterized via genome-wide experimentation, pose a challenge for integration and analytic efforts [68]. These challenges may be mitigated via combinatorial integration of fundamental data attributes into generalized data structures that can be mined for patterns or emergent gene-disease relationships. GeneWeaver [69] for example, relies on a bi-partite data model [70] and heterogeneous data networks [71] to integrate and analyze functional genomics data such as differential expression studies, GWAS, curated annotations, and QTL mapping studies through a single data structure that facilitates

aggregation of information. Harmonizome [72], on the other hand, aggregates functional genomics studies from a variety of sources by implementing an association matrix across shared attributes and relying on machine learning approaches to identify novel patterns.

Fundamental integration through knowledge graphs may also be applied to large scale heterogenous analysis. KnowEng [73] uses a knowledge network to navigate the integration of statistical experimental data and contextualized user information to identify human and mouse interactions. Aggregated knowledge networks can be analyzed using traditional network mining approaches or machine learning. Other tools, such as HumanBase [74] or the DIAMOnD [75] algorithm, also takes advantage of traversing large *ad hoc* networks of functional connectivity. Networks are navigated through machine learning or association matrices to connect multi-species gene or variant relationships.

There are many approaches to cross-species comparative genomics and phenomics integration (e.g. Table 2) and analysis must optimize among competing needs of computing scalability, data accessibility, and data scope. For example, the sheer number of variants in humans and rodents and the unbounded phenotype dimension lead to the problem of phenomenal computational scale. The tremendous heterogeneity of model organism data sets, from mutation characterization studies, curated pathway and gene annotation sets, and extensive genetic and genomic data at the level of genes and variants, presents a problem of size, scope and complexity, in the realm of Big Data problems, requiring computationally scalable solutions.

BIG DATA AND THE INTEGRATION OF HUMAN AND MODEL ORGANISM STUDIES IN PSYCHIATRIC GENETICS

Cross species analysis typically happens at the level of abstracted relations among variants or genes and can thus be quite reduced in scale. However, 1) the scope of genomic studies is completely unbounded and it is possible to find hundreds, if not thousands of animal studies of disease relevant neurobiology and 2) the parsing and representation of genomic variants from diverse data sources and their mappings onto one another does not scale so easily. Retaining this traceable mapping while allowing integrative and interactive analysis is a problem of high complexity and scale. The storage, analysis, distribution and integration of human and model organism functional genomic data are especially challenging, as they embody typical problems encountered in the Big Data world [76] often referred to as the four V's of data--Volume, Variety, Velocity, and Veracity.

The sheer volume of data required to support comprehensive cross-species data integration of genes and individual variants is staggering. For example, if we assume that the average number of coding genes in mammalian genomes is approximately 25,000, then constructing rudimentary connections among the genes in five species would produce ½n(n-1) relationships, where n is the number of genes in the network. If represented as a graph, with each edge representing a relationship, the graph would be enormous but tractable, comprising ~7.8E9 edges. But, the genome is only one dimension of the problem. The other is the sheer number of contexts in which that genome is experimentally profiled. With thousands of human and model organism addiction genomics data sets, and hundreds of thousands of species-specific pathway data, brain regional transcriptomes and other relevant data resources, one quickly reaches a problem requiring scalable solutions. Analysis of a handful of organisms can therefore be handled with large, conventional high-performance computing systems. At the variant level, however, the relationship problem is greatly compounded. Known variants, which outnumber

genes within the typical model organisms by more than 20,000 to 1, would naïvely require ~1.25E17 edge relationships. While intelligent approaches for computing on large graphs, such as taking advantage of partitioning [77], sparse connectivity [78], or heuristics [79], can aid in the management and analysis of these relationships, exhaustive examination of static graphs of this potential size is intractable due to computing limitations, storage and real-time accessibility. As the number of genomic experiments continue to grow, particularly in the model organism space, one viable option may be the dynamic analysis of data sets using elastic on-demand cloud services that make use of horizontally-scalable computing to efficiently distribute computing tasks to address very specific questions.

A corollary to the volume/variety of data associated with variant mapping across species is the velocity at which it is produced, and, subsequently, the rate at which it must be collated, curated, and made accessible. With over 4500 eukaryotic genomes assembled over the last decade [80], it has been argued that genome-scale data will be bigger than Big Data associated with astronomy, YouTube, and Twitter by 2025 [76]. To complicate the processes used to integrate the vast scope of data are data sharing policies that historically do not require automated sharing of model organism data, resulting in data analysis processes that result primarily from *ad hoc* relationships [81]. To mitigate the stresses imposed by data velocity, it is critical to devise a means to access, integrate, and dynamically update these data in a manner that avoids redundancies and keeps data provenance intact. While it is inevitable that there will be an uneven integration of data from a variety of sources, it is incumbent on the bioinformatics community to create systems to rapidly track intentional methodologies for data cleaning and reduction through the discovery of duplicated or deprecated data.

By addressing these problems in Big Data, scalable applications in integrative functional genomics for psychiatric genomics are enabled (Figure 2). The integrated, global mapping of trait regulatory variants across species through target genes can facilitate the integration of model organism genomic data to fill the mechanistic knowledge gap between non-coding human genetic variant and human disease. This integration can be accomplished through the aggregation of curated and high-throughput experimental data from multiple domain-specific resources. Data resources such as GTEx [29], ENCODE [82], and Roadmap Epigenomics [83], provide extensive coverage of genomic regulatory features and gene-regulatory mechanisms. High-level regulatory features including CTCF binding sites, enhancers, open chromatin, promoter, promoter flanking, and transcription factor binding site attributes can all be retrieved from regulation databases [84]. These features can be annotated to genomic variants from the Ensembl variation database [85], for example, to identify regulatory variants within regions of interest. Identifying putative regulatory interactions between regulatory variants and genes can be accomplished through layering several approaches. Topologically associated domains (TAD), verified from Hi-C studies and integrated from published studies and the ENCODE resource, can be used to delineate putative gene-regulatory boundaries and all combinations of regulatory variants and genes that are associated within the boundary. Experimentally confirmed feature-gene interactions mediated by RNA polymerase II (RNAPII) and identified using ChIA-PET studies, sourced from ENCODE and various publications, can also be used. Finally, eQTLs can identify variant influences on specific genes.

Compounding the issues encountered by the complexity of raw data is the potential for underlying data bias and the subsequent difficulty of attributing veracity to the data. There is an implicit bias in the sampling of genes represented in an experimentally derived genomic data set

because each genomic technology and especially a curated genomic data resource is based on a different breadth, e.g. individual mutation studies curated from literature by the Model Organism Databases vs. genome-wide gene expression by RNA-seq data. Differing approaches affect the rate of false positives in the data set. For example, QTL positional candidate sets may have many genes with likely only one or a few true positives, in contrast to differential gene expression sets for which the statistical threshold defines a false discovery rate. Semi-quantitative or quantitative scores for these data sets need to be created to reduce our reliance on qualitative scoring. Enrichment analyses and systems genetic correlation tools suffer from annotation bias in that one often retrieves results representing areas of investigation that are dense with information, resulting in apparent patterns and trends that are an artefact of coverage. Data resources like GTEx also suffer from biases based on uneven sample size, and the particular tissues and conditions investigated. The net effect of the uneven statistical power in these data resources is to upwardly bias well-powered but less relevant findings, in which tissues or phenotypes are spuriously associated with disease. Therefore, it is important to consider error-rate controls, and other procedures, but also the uniformity of analysis in the data used in analysis.

Multi-tissue eQTL data can be integrated to provide context-specific variant mapping. Primarily derived from the GTEx project or model organism resources such as GeneNetwork [30], data from mouse, rat and human genetics experiments represent a diverse and deep pool of data. Single cell RNA (scRNA) enables the exciting possibility to investigate eQTLs and gene coexpression in complex, multicellular tissues. For example, scRNA sequences have been used to create high fidelity classifications of brain regions based on local variants [86]. Furthermore, scRNA have been used to identify cell type-specific cis-eQTLs and variant co-expression networks [87]. Gene expression genetics studies in model organisms have tremendous

precision with new populations like the Diversity Outbred segregating 45 million mouse genetic variants comprising 90% of the known mouse genetic variome [88]. Recombinations are at extremely high precision, and large mapping population sample sizes for an increasing number of brain regions, and the derivation of this population from eight founder strains provides a means of reducing eQTLs to a small handful of regulatory variants at the SNP level [89]. As such, it is possible to identify eQTL variants which may affect one of several gene regulatory mechanisms targeting a human orthologue, and to assess its effect on mouse phenomics, cellular gene expression, or other endpoints *in silico*, *in vitro*, or *in vivo*. Many of these tools provide browser-based and limited scriptable interfaces with continued adoption of new technologies, but exposing model organism eQTL data to large-scale dynamic tools for graphical integration would be of tremendous utility in readily enabling facile interrogation of variant-gene relations.

MULTIPLE APPLICATIONS ARE READILY POSSIBLE WITH INTEGRATED DATA STRUCTURES

A compelling approach to the prioritization of GWAS variants enabled by BigData integration is the use integrated cross-species data to identify and characterize those variants with a known mechanistic role in neurobiological pathways to disease, or to identify human variants with highly specific hypothesized roles in particular cases of disease, such as the widely studied ADH1B in AUDs. Although current applications and analytic implementations do not fully take advantage of large scale data resources, the emerging scale of data and high-volume comparative analyses will most certainly merit scalable approaches in the near future. Most present approaches do not yet harness the full capacity of cross-species comparative analyses at scale, and initial applications have been necessarily focused on small, single locus problems. However,

these simple applications are ripe for extrapolation to global questions about the neurobiological mechanisms of addiction. One promising application of multi-species epigenomic integration is comparative gene regulation. Now that characterization of gene-regulatory components (e.g., enhancers, TF binding sites) and their putative gene targets is improving, integrative methods can identify shared genomic regulators across species. In one example, from studies on alcohol dependence and cholesterol, at least 4,000 SNPs from human GWAS can be mapped onto the mouse genome [90]. Furthermore, some of these SNPs, which are involved in human liver function, can be mapped to liver-specific enhancers in mice [91]. This type of comparative analysis could be used to identify convergent regulatory features and variants across species, enabling the development of mouse models for testing SNP causality in humans. Integrative systems have successfully been used to identify disease-relevant genes and to identify gene-regulatory SNPs involved in alcohol preference and withdrawal involved in epigenetic regulation in mice at a distal enhancer element [92]. Query of public genetic data resources indicates that variation in the same gene occurs in humans, via a promoter variant, rather than an enhancer [93].

Several recent approaches have been developed for prioritization of disease relevant genes and variants from integrative omics analysis. These tools utilize large integration pipelines coupled to networking and statistical tools to establish a relative importance (e.g., priority indexing) of variants across tissues of interest focusing on immune-mediated traits. For example, Wang et al., develop a risk gene selection method, called iRIGs [18], which incorporates GWAS and a number of genomic features including expression, chromatin interactions, and gene-regulatory data into a Bayesian framework for prioritization. This framework prioritizes genes within a small 2MB region near risk loci identified from GWAS using a select set of epigenetics

including promoter, enhancer, and chromatin interactions from Hi-C studies. A similar approach, developed by Fang et al., utilizes a priority index (Pi) pipeline [94] designed to prioritize genes from GWAS variants for specific immune traits. Pi combines genomic predictors in the form of gene proximities, chromatin interactions, and expression modulation evidence (eQTLs) with network-based models to prioritize trait-gene associations. To date, these approaches have not been applied to model organism data, but they most certainly can be. Furthermore, with the implementation of cross species variant mapping such as those presented in Figure 1, they can exploit the broad, heterogeneous multi-species data corpus.

Another application is to compare sets of trait associated human and model organism genomic data to identify similarly regulated disease-relevant traits suitable for convergent validation experiments. Mapping of human disease–related characteristics onto model organism behaviors has been a controversial area of research, and for many, the perceived relevance of animal models is hindered by ever-refined definitions of face validity [95]. This argument misses the point that a model is by definition a simplification of a system that renders it amenable to particular types of study, including validation. Animal models, themselves, have successfully been used to measure the efficacy of drugs and validate various drug targets [51]. Further, there may be sufficient consilience between human disease traits (such as the various aspects of alcohol use disorders) and those modeled in animals (e.g., ethanol intake) at a genomic level (rg=0.77 between problem drinking and typical alcohol intake [3]) to allow for careful cross-species data integration for these sub-facets of human disease. Research targeting behavioral mechanisms that do converge across species does not discount or diminish the need to study the remaining complexity in the human phenotype. Rather, it serves as a powerful means of

discovery of the nature of vulnerability and resilience to those components of psychiatric disorders that, in their many manifestations and potentially relevant classifications, are amenable to biological insights, and thus, promising targets for therapeutic discovery.

Finally, the prioritization of variants for use in polygenic risk analysis can be refined. Savvy integrative methods can be combined to achieve sets of variants that meaningfully contribute to trait variation from a broad network of genes. Aggregating across the tools and databases listed in the current review will help researchers to match 1) variants to genes, 2) genes to biological functions, 3) functions to plausible molecular mechanisms—ultimately achieving more robust effects with high signal-to-noise ratios—and 4) traits and disease characteristics within and across species. A few studies [96] have constructed polygenic scores from variants in genes known for disease pathology or targets co-expressed with putative trait genes from relevant brain tissues (via GeneNetwork), both of which demonstrated increased prediction than a random sets of genes and achieved trait specificity in mice [96] and humans [97]. But not all biologically informed polygenic scores exhibit significant prediction [98] and these methods have not been benchmarked with classical approaches selecting specific statistical criterion (e.g. p-value threshold; PRSice [99]) nor approaches that combine both statistical and alternative biological information (e.g., LD; LDpred [100], PleioPRED [101], AnnoPred [21]). A mixture of these techniques is likely required to best inform gene and variant prioritization in human GWAS studies.

FUTURE RESEARCH DIRECTIONS

The multiple strategies we have outlined can be used to address the challenges and opportunities for the integration of diverse model organism data sets to augment the interpretation of GWAS and define genes and molecular pathways that underlie aspects of psychiatrically relevant phenotypes. Heterogeneous functional genomics leverages the combined information in population genetic diversity, systems biology, gene regulatory analysis, and advanced phenotypic measurements to identify and characterize mechanisms of psychiatric disorders of the greatest complexity. Much work remains to facilitate dynamic data integration across these data types. The continued generation of adequately powered and broadly unbiased data resources in neurogenomics is essential across multiple species. Data sharing policies and practices along with platforms for data sharing and data integration are required. Community standards and practices that make data Findable, Accesible, Interoperable and Reproducible (FAIR) need to be adopted and resourced so that all researchers engaged in the generation and analysis of integrative functional genomics data have the capability of contributing to and benefiting from data integration. Development of analytic approaches and algorithms are also required for diverse applications in functional genomic data integration. Scalable computational solutions that allow for such high dimensional data integration will enable a growing array of tools and approaches for the discovery of unknown mechanisms underlying psychiatric disorders, providing a more complete understanding of disease mechanisms.

FUNDING AND DISCLOSURE

The authors declare that they do not have any conflicts of interest related to the content of the paper. AA and ECJ receive support from NIH MH109532; DA32573 (AA); F32AA027435

(ECJ). EJC, TR, EJB and JAB receive support from NIH AA018776; and EJC from P50 DA039841 (EJC). RHCP, JAB, and SH receive support from NIH DA042103 (RHCP).

ACKNOWLEDGEMENTS

The Authors thank Stephen Krasinski of The Jackson Laboratory for assistance with this manuscript.

AUTHOR CONTRIBUTIONS

EJC conceived of the overall manuscript and oversaw its preparation. All authors contributed to the content and writing of the Manuscript. EJC, EJB, and TR conceived of Figure 1, prepared by TR. JAB researched and prepared Table 1.

REFERENCES

[1] O. B. Smeland, O. Frei, C.-C. Fan, A. Shadrin, A. M. Dale, and O. A. Andreassen, "The emerging pattern of shared polygenic architecture of psychiatric disorders, conceptual and methodological challenges," *Psychiatr. Genet.*, vol. 29, no. 5, pp. 152–159, 2019, doi: 10.1097/YPG.0000000000000234.

[2] Schizophrenia Working Group of the Psychiatric Genomics Consortium, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, pp. 421–427, Jul. 2014, doi: 10.1038/nature13595.

[3] H. Zhou *et al.*, "Meta-analysis of problematic alcohol use in 435,563 individuals identifies 29 risk variants and yields insights into biology, pleiotropy and causality," *bioRxiv*, p. 738088, Aug. 2019, doi: 10.1101/738088.

[4] D. M. Howard *et al.*, "Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions," *Nat Neurosci*, vol. 22, no. 3, Art. no. 3, Mar. 2019, doi: 10.1038/s41593-018-0326-7.

[5] Cross-Disorder Group of the Psychiatric Genomics Consortium. Electronic address: plee0@mgh.harvard.edu and Cross-Disorder Group of the Psychiatric Genomics Consortium, "Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders," *Cell*, vol. 179, no. 7, pp. 1469-1482.e11, Dec. 2019, doi: 10.1016/j.cell.2019.11.020.

[6] M. Liu *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.*, vol. 51, no. 2, pp. 237–244, 2019, doi: 10.1038/s41588-018-0307-5.
[7] S. Sanchez-Roige *et al.*, "Genome-Wide Association Studies of Impulsive Personality Traits (BIS-11 and UPPS-P) and Drug Experimentation in up to 22,861 Adult Research Participants Identify Loci in the CACNA1I and CADM2 genes," *J. Neurosci.*, vol. 39, no. 13, pp. 2562–2572, Mar. 2019, doi: 10.1523/JNEUROSCI.2662-18.2019.
[8] H. R. Kranzler *et al.*, "Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations," *Nat Commun*, vol. 10, no. 1, p. 1499, 02 2019, doi: 10.1038/s41467-019-09480-8.
[9] M. D. Gallagher and A. S. Chen-Plotkin, "The Post-GWAS Era: From Association to Function," *Am. J. Hum. Genet.*, vol. 102, no. 5, pp. 717–730, 03 2018, doi: 10.1016/j.ajhg.2018.04.002.
[10] P. F. Sullivan and D. H. Geschwind, "Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders," *Cell*, vol. 177, no. 1, pp. 162–183, Mar. 2019, doi: 10.1016/j.cell.2019.01.015.
[11] M. Wang and S. Xu, "Statistical power in genome-wide association studies and quantitative trait locus mapping," *Heredity (Edinb)*, vol. 123, no. 3, pp. 287–306, 2019, doi: 10.1038/s41437-019-0205-3.
[12] F. Dudbridge, "Power and Predictive Accuracy of Polygenic Risk Scores," *PLoS Genetics*, vol. 9, no. 3, p. e1003348, Mar. 2013, doi: 10.1371/journal.pgen.1003348.
[13] P. Wainschtein *et al.*, "Recovery of trait heritability from whole genome sequence data," *bioRxiv*, p. 588020, Mar. 2019, doi: 10.1101/588020.
[14] A. Caspi *et al.*, "The p Factor: One General Psychopathology Factor in the Structure of Psychiatric Disorders?," *Clin Psychol Sci*, vol. 2, no. 2, pp. 119–137, Mar. 2014, doi: 10.1177/2167702613497473.
[15] K. Watanabe, E. Taskesen, A. van Bochoven, and D. Posthuma, "Functional mapping and annotation of genetic associations with FUMA," *Nat Commun*, vol. 8, no. 1, p. 1826, 28 2017, doi: 10.1038/s41467-017-01261-5.
[16] S. Ning *et al.*, "LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs," *BMC Bioinformatics*, vol. 15, p. 152, May 2014, doi: 10.1186/1471-2105-15-152.
[17] Y. Wu *et al.*, "Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits," *Nature Communications*, vol. 9, no. 1, Art. no. 1, Mar. 2018, doi: 10.1038/s41467-018-03371-0.
[18] Q. Wang *et al.*, "A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data," *Nat. Neurosci.*, vol. 22, no. 5, pp. 691–699, 2019, doi: 10.1038/s41593-019-0382-7.
[19] G. Kichaev *et al.*, "Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies," *PLOS Genetics*, vol. 10, no. 10, p. e1004722, Oct. 2014, doi: 10.1371/journal.pgen.1004722.
[20] N. Mancuso *et al.*, "Probabilistic fine-mapping of transcriptome-wide association studies," *Nature Genetics*, vol. 51, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s41588-019-0367-1.
[21] Y. Hu *et al.*, "Leveraging functional annotations in genetic risk prediction for human complex diseases," *PLOS Computational Biology*, vol. 13, no. 6, p. e1005589, Jun. 2017, doi: 10.1371/journal.pcbi.1005589.

[22] C. Márquez-Luna *et al.*, "LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets," *bioRxiv*, p. 375337, May 2020, doi: 10.1101/375337.
[23] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics Data Integration, Interpretation, and Its Application," *Bioinform Biol Insights*, vol. 14, p. 1177932219899051, 2020, doi: 10.1177/1177932219899051.
[24] G. Kichaev *et al.*, "Leveraging Polygenic Functional Enrichment to Improve GWAS Power," *The American Journal of Human Genetics*, vol. 104, no. 1, pp. 65–75, Jan. 2019, doi: 10.1016/j.ajhg.2018.11.008.
[25] R. H. C. Palmer *et al.*, "Cross-Species Integration of Transcriptomic Effects of Tobacco and Nicotine Exposure Helps to Prioritize Genetic Effects on Human Tobacco Consumption," *bioRxiv*, p. 2019.12.23.887083, Dec. 2019, doi: 10.1101/2019.12.23.887083.
[26] K. M. Mignogna, S. A. Bacanu, B. P. Riley, A. R. Wolen, and M. F. Miles, "Cross-species alcohol dependence-associated gene networks: Co-analysis of mouse brain gene expression and human genome-wide association data," *PLOS ONE*, vol. 14, no. 4, p. e0202063, Apr. 2019, doi: 10.1371/journal.pone.0202063.
[27] E. J. Chesler *et al.*, "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function," *Nat. Genet.*, vol. 37, no. 3, pp. 233–242, Mar. 2005, doi: 10.1038/ng1518.
[28] J. C. Crabbe, "Progress With Nonhuman Animal Models of Addiction," *J Stud Alcohol Drugs*, vol. 77, no. 5, pp. 696–699, 2016, doi: 10.15288/jsad.2016.77.696.
[29] GTEx Consortium, "The Genotype-Tissue Expression (GTEx) project," *Nat. Genet.*, vol. 45, no. 6, pp. 580–585, Jun. 2013, doi: 10.1038/ng.2653.
[30] M. K. Mulligan, K. Mozhui, P. Prins, and R. W. Williams, "GeneNetwork: A Toolbox for Systems Genetics," *Methods Mol. Biol.*, vol. 1488, pp. 75–120, 2017, doi: 10.1007/978-1-4939-6427-7_4.
[31] PsychENCODE Consortium *et al.*, "The PsychENCODE project," *Nat. Neurosci.*, vol. 18, no. 12, pp. 1707–1712, Dec. 2015, doi: 10.1038/nn.4156.
[32] S. E. Celniker *et al.*, "Unlocking the secrets of the genome," *Nature*, vol. 459, no. 7249, pp. 927–930, Jun. 2009, doi: 10.1038/459927a.
[33] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002, doi: 10.1093/nar/30.1.207.
[34] M. E. Dickinson *et al.*, "High-throughput discovery of novel developmental phenotypes," *Nature*, vol. 537, no. 7621, pp. 508–514, 22 2016, doi: 10.1038/nature19356.
[35] C. Durrant *et al.*, "Bioinformatics tools and database resources for systems genetics analysis in mice--a short review and an evaluation of future needs," *Brief. Bioinformatics*, vol. 13, no. 2, pp. 135–142, Mar. 2012, doi: 10.1093/bib/bbr026.
[36] M. A. Bogue *et al.*, "Mouse Phenome Database: a data repository and analysis suite for curated primary mouse phenotype data," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D716–D723, 08 2020, doi: 10.1093/nar/gkz1032.
[37] M. A. Bogue *et al.*, "Mouse Phenome Database: an integrative database and analysis suite for curated empirical phenotype data from laboratory mice," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D843–D850, 04 2018, doi: 10.1093/nar/gkx1082.
[38] M. Nishiguchi *et al.*, "Different blood acetaldehyde concentration following ethanol administration in a newly developed high alcohol preference and low alcohol preference rat

model system," *Alcohol Alcohol.*, vol. 37, no. 1, pp. 9–12, Feb. 2002, doi: 10.1093/alcalc/37.1.9.

[39] B. Oberlin, C. Best, L. Matson, A. Henderson, and N. Grahame, "Derivation and characterization of replicate high- and low-alcohol preferring lines of mice and a high-drinking crossed HAP line," *Behav. Genet.*, vol. 41, no. 2, pp. 288–302, Mar. 2011, doi: 10.1007/s10519-010-9394-5.

[40] S. E. Bergeson, R. Kyle Warren, J. C. Crabbe, P. Metten, V. Gene Erwin, and J. K. Belknap, "Chromosomal loci influencing chronic alcohol withdrawal severity," *Mamm. Genome*, vol. 14, no. 7, pp. 454–463, Jul. 2003, doi: 10.1007/s00335-002-2254-4.

[41] A. E. Adkins *et al.*, "Genomewide Association Study of Alcohol Dependence Identifies Risk Loci Altering Ethanol-response Behaviors in Model Organisms," *Alcohol Clin Exp Res*, vol. 41, no. 5, pp. 911–928, May 2017, doi: 10.1111/acer.13362.

[42] G. Schumann *et al.*, "KLB is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference," *Proc Natl Acad Sci U S A*, vol. 113, no. 50, pp. 14372–14377, Dec. 2016, doi: 10.1073/pnas.1611243113.

[43] A. Sekar *et al.*, "Schizophrenia risk from complex variation of complement component 4," *Nature*, vol. 530, no. 7589, pp. 177–183, Feb. 2016, doi: 10.1038/nature16549.

[44] S. Rangaraju *et al.*, "Identification and therapeutic modulation of a pro-inflammatory subset of disease-associated-microglia in Alzheimer's disease," *Mol Neurodegener*, vol. 13, no. 1, p. 24, 21 2018, doi: 10.1186/s13024-018-0254-8.

[45] A. R. Wolen *et al.*, "Genetic dissection of acute ethanol responsive gene networks in prefrontal cortex: functional and mechanistic implications," *PLoS ONE*, vol. 7, no. 4, p. e33575, 2012, doi: 10.1371/journal.pone.0033575.

[46] P. Turley *et al.*, "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nat. Genet.*, vol. 50, no. 2, pp. 229–237, 2018, doi: 10.1038/s41588-017-0009-4.

[47] S. Sanchez-Roige *et al.*, "Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts," *Am J Psychiatry*, vol. 176, no. 2, pp. 107–118, Feb. 2019, doi: 10.1176/appi.ajp.2018.18040369.

[48] J. C. Crabbe, "Translational behaviour-genetic studies of alcohol: are we there yet?," *Genes Brain Behav.*, vol. 11, no. 4, pp. 375–386, Jun. 2012, doi: 10.1111/j.1601-183X.2012.00798.x.

[49] H.-L. Zhang *et al.*, "Comparative analysis of cellular expression pattern of schizophrenia risk genes in human versus mouse cortex," *Cell Biosci*, vol. 9, p. 89, 2019, doi: 10.1186/s13578-019-0352-5.

[50] M. P. van den Heuvel *et al.*, "Evolutionary modifications in human brain connectivity associated with schizophrenia," *Brain*, vol. 142, no. 12, pp. 3991–4002, Dec. 2019, doi: 10.1093/brain/awz330.

[51] V. Krishnan and E. J. Nestler, "Animal Models of Depression: Molecular Perspectives," *Curr Top Behav Neurosci*, vol. 7, pp. 121–147, 2011, doi: 10.1007/7854_2010_108.

[52] E. J. Baker *et al.*, "Ontological Discovery Environment: a system for integrating gene-phenotype associations," *Genomics*, vol. 94, no. 6, pp. 377–387, Dec. 2009, doi: 10.1016/j.ygeno.2009.08.016.

[53] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, no. 1, p. 122, 06 2016, doi: 10.1186/s13059-016-0974-4.

[54] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, Sep. 2010, doi: 10.1093/nar/gkq603.
[55] A. Brodie, J. R. Azaria, and Y. Ofran, "How far from the SNP may the causative genes be?," *Nucleic Acids Res*, vol. 44, no. 13, pp. 6046–6054, Jul. 2016, doi: 10.1093/nar/gkw500.
[56] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma, "MAGMA: generalized gene-set analysis of GWAS data," *PLoS Comput. Biol.*, vol. 11, no. 4, p. e1004219, Apr. 2015, doi: 10.1371/journal.pcbi.1004219.
[57] A. N. Barbeira *et al.*, "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics," *Nat Commun*, vol. 9, no. 1, Art. no. 1, May 2018, doi: 10.1038/s41467-018-03621-1.
[58] A. Gusev *et al.*, "Integrative approaches for large-scale transcriptome-wide association studies," *Nat. Genet.*, vol. 48, no. 3, pp. 245–252, Mar. 2016, doi: 10.1038/ng.3506.
[59] N. Y. A. Sey, H. Fauni, W. Ma, and H. Won, "Connecting gene regulatory relationships to neurobiological mechanisms of brain disorders," *bioRxiv*, p. 681353, Jun. 2019, doi: 10.1101/681353.
[60] Alliance of Genome Resources Consortium, "Alliance of Genome Resources Portal: unified model organism research platform," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D650–D658, Jan. 2020, doi: 10.1093/nar/gkz813.
[61] E. L. L. Sonnhammer and G. Östlund, "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic," *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D234-239, Jan. 2015, doi: 10.1093/nar/gku1203.
[62] U. M. Marigorta, J. A. Rodríguez, G. Gibson, and A. Navarro, "Replicability and Prediction: Lessons and Challenges from GWAS," *Trends Genet.*, vol. 34, no. 7, pp. 504–517, 2018, doi: 10.1016/j.tig.2018.03.005.
[63] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, "Linking human diseases to animal models using ontology-based phenotype annotation," *PLoS Biol.*, vol. 7, no. 11, p. e1000247, Nov. 2009, doi: 10.1371/journal.pbio.1000247.
[64] C. J. Mungall *et al.*, "The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D712–D722, 2017, doi: 10.1093/nar/gkw1128.
[65] D. Smedley *et al.*, "PhenoDigm: analyzing curated annotations to associate animal models with human diseases," *Database (Oxford)*, vol. 2013, p. bat025, 2013, doi: 10.1093/database/bat025.
[66] W. P. Bone *et al.*, "Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency," *Genet. Med.*, vol. 18, no. 6, pp. 608–617, 2016, doi: 10.1038/gim.2015.137.
[67] P. N. Robinson *et al.*, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Res.*, vol. 24, no. 2, pp. 340–348, Feb. 2014, doi: 10.1101/gr.160325.113.
[68] J. A. Bubier, C. A. Phillips, M. A. Langston, E. J. Baker, and E. J. Chesler, "GeneWeaver: finding consilience in heterogeneous cross-species functional genomics data," *Mamm. Genome*, vol. 26, no. 9–10, pp. 556–566, Oct. 2015, doi: 10.1007/s00335-015-9575-x.

[69] E. Baker, J. A. Bubier, T. Reynolds, M. A. Langston, and E. J. Chesler, "GeneWeaver: data driven alignment of cross-species genomics in biology and disease," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D555-559, Jan. 2016, doi: 10.1093/nar/gkv1329.
[70] E. J. Baker, J. J. Jay, J. A. Bubier, M. A. Langston, and E. J. Chesler, "GeneWeaver: a web-based system for integrative functional genomics," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1067-1076, Jan. 2012, doi: 10.1093/nar/gkr968.
[71] T. Reynolds, J. A. Bubier, M. A. Langston, E. J. Chesler, and E. J. Baker, "Finding human gene-disease associations using a Network Enhanced Similarity Search (NESS) of multi-species heterogeneous functional genomics data," *bioRxiv*, p. 2020.03.11.987552, Mar. 2020, doi: 10.1101/2020.03.11.987552.
[72] A. D. Rouillard *et al.*, "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins," *Database (Oxford)*, vol. 2016, 2016, doi: 10.1093/database/baw100.
[73] S. Sinha, J. Song, R. Weinshilboum, V. Jongeneel, and J. Han, "KnowEnG: a knowledge engine for genomics," *J Am Med Inform Assoc*, vol. 22, no. 6, pp. 1115–1119, Nov. 2015, doi: 10.1093/jamia/ocv090.
[74] C. S. Greene *et al.*, "Understanding multicellular function and disease with human tissue-specific networks," *Nat Genet*, vol. 47, no. 6, pp. 569–576, Jun. 2015, doi: 10.1038/ng.3259.
[75] S. D. Ghiassian, J. Menche, and A.-L. Barabási, "A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome," *PLoS Comput. Biol.*, vol. 11, no. 4, p. e1004120, Apr. 2015, doi: 10.1371/journal.pcbi.1004120.
[76] Z. D. Stephens *et al.*, "Big Data: Astronomical or Genomical?," *PLOS Biology*, vol. 13, no. 7, p. e1002195, Jul. 2015, doi: 10.1371/journal.pbio.1002195.
[77] E. G. Boman, K. D. Devine, and S. Rajamanickam, "Scalable matrix computations on large scale-free graphs using 2D graph partitioning," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, Denver, Colorado, Nov. 2013, pp. 1–12, doi: 10.1145/2503210.2503293.
[78] M. Latapy, "Main-memory triangle computations for very large (sparse (power-law)) graphs," *Theoretical Computer Science*, vol. 407, no. 1, pp. 458–473, Nov. 2008, doi: 10.1016/j.tcs.2008.07.017.
[79] I. Stanton and G. Kliot, "Streaming graph partitioning for large distributed graphs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Beijing, China, Aug. 2012, pp. 1222–1230, doi: 10.1145/2339530.2339722.
[80] D. A. Benson *et al.*, "GenBank," *Nucleic Acids Res*, vol. 46, no. Database issue, pp. D41–D47, Jan. 2018, doi: 10.1093/nar/gkx1094.
[81] D. Field *et al.*, "Megascience. 'Omics data sharing," *Science*, vol. 326, no. 5950, pp. 234–236, Oct. 2009, doi: 10.1126/science.1180598.
[82] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.
[83] B. E. Bernstein *et al.*, "The NIH Roadmap Epigenomics Mapping Consortium," *Nat. Biotechnol.*, vol. 28, no. 10, pp. 1045–1048, Oct. 2010, doi: 10.1038/nbt1010-1045.
[84] D. R. Zerbino *et al.*, "Ensembl regulation resources," *Database (Oxford)*, vol. 2016, 2016, doi: 10.1093/database/bav119.

[85] S. E. Hunt *et al.*, "Ensembl variation resources," *Database (Oxford)*, vol. 2018, 01 2018, doi: 10.1093/database/bay119.
[86] J.-F. Poulin, B. Tasic, J. Hjerling-Leffler, J. M. Trimarchi, and R. Awatramani, "Disentangling neural cell diversity using single-cell transcriptomics," *Nature Neuroscience*, vol. 19, no. 9, Art. no. 9, Sep. 2016, doi: 10.1038/nn.4366.
[87] M. G. P. van der Wijst, H. Brugge, D. H. de Vries, P. Deelen, M. A. Swertz, and L. Franke, "Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs," *Nature Genetics*, vol. 50, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/s41588-018-0089-9.
[88] A. Roberts, F. Pardo-Manuel de Villena, W. Wang, L. McMillan, and D. W. Threadgill, "The polymorphism architecture of mouse genetic resources elucidated using genome-wide resequencing data: implications for QTL discovery and systems genetics," *Mamm. Genome*, vol. 18, no. 6–7, pp. 473–481, Jul. 2007, doi: 10.1007/s00335-007-9045-1.
[89] D. A. Skelly, N. Raghupathy, R. F. Robledo, J. H. Graber, and E. J. Chesler, "Reference Trait Analysis Reveals Correlations Between Gene Expression and Quantitative Traits in Disjoint Samples," *Genetics*, vol. 212, no. 3, pp. 919–929, 2019, doi: 10.1534/genetics.118.301865.
[90] F. Yue *et al.*, "A comparative encyclopedia of DNA elements in the mouse genome," *Nature*, vol. 515, no. 7527, pp. 355–364, Nov. 2014, doi: 10.1038/nature13992.
[91] A. Breschi, T. R. Gingeras, and R. Guigó, "Comparative transcriptomics in human and mouse," *Nat. Rev. Genet.*, vol. 18, no. 7, pp. 425–440, 2017, doi: 10.1038/nrg.2017.19.
[92] J. A. Bubier *et al.*, "Identification of a QTL in Mus musculus for Alcohol Preference, Withdrawal, and Ap3m2 Expression Using Integrative Functional Genomics and Precision," *Genetics*, vol. 197, no. 4, pp. 1377–1393, Aug. 2014, doi: 10.1534/genetics.114.166165.
[93] GTEx Consortium, "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans," *Science*, vol. 348, no. 6235, pp. 648–660, May 2015, doi: 10.1126/science.1262110.
[94] H. Fang *et al.*, "A genetics-led approach defines the drug target landscape of 30 immune-related traits," *Nat. Genet.*, vol. 51, no. 7, pp. 1082–1091, 2019, doi: 10.1038/s41588-019-0456-1.
[95] E. J. Nestler and S. E. Hyman, "Animal models of neuropsychiatric disorders," *Nature Neuroscience*, vol. 13, no. 10, pp. 1161–1169, Oct. 2010, doi: 10.1038/nn.2647.
[96] S. M. Neuner, S. E. Heuer, M. J. Huentelman, K. M. S. O'Connell, and C. C. Kaczorowski, "Harnessing Genetic Complexity to Enhance Translatability of Alzheimer's Disease Mouse Models: A Path toward Precision Medicine," *Neuron*, vol. 101, no. 3, pp. 399-411.e5, 06 2019, doi: 10.1016/j.neuron.2018.11.040.
[97] S. A. Hari Dass *et al.*, "A biologically-informed polygenic score identifies endophenotypes and clinical conditions associated with the insulin receptor function on specific brain regions," *EBioMedicine*, vol. 42, pp. 188–202, Apr. 2019, doi: 10.1016/j.ebiom.2019.03.051.
[98] S. Van der Auwera *et al.*, "Predicting brain structure in population-based samples with biologically informed genetic scores for schizophrenia," *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, vol. 174, no. 3, pp. 324–332, Apr. 2017, doi: 10.1002/ajmg.b.32519.

[99] J. Euesden, C. M. Lewis, and P. F. O'Reilly, "PRSice: Polygenic Risk Score software," *Bioinformatics*, vol. 31, no. 9, pp. 1466–1468, May 2015, doi: 10.1093/bioinformatics/btu848.

[100] B. J. Vilhjálmsson *et al.*, "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores," *Am. J. Hum. Genet.*, vol. 97, no. 4, pp. 576–592, Oct. 2015, doi: 10.1016/j.ajhg.2015.09.001.

[101] Y. Hu, Q. Lu, W. Liu, Y. Zhang, M. Li, and H. Zhao, "Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction," *PLoS Genet.*, vol. 13, no. 6, p. e1006836, Jun. 2017, doi: 10.1371/journal.pgen.1006836.

Figure Legends:

Fig 1. Multi-species genomic and epigenomic data integration. Genetic variation, gene regulation, and homology datasets are retrieved from a variety of publicly available resources and data repositories. Human (V_H) and mouse (V_M) variants are connected to the gene (G_M, G_H) that either contains a coding variant or is regulated by a non-coding variant. Epigenetic markers and regulatory features (R_M, R_H) are retrieved from ENCODE and Ensembl, then overlapped with genetic variation data from Ensembl and NCBI in order to identify regulatory variants (V_M, V_H). Regulatory variants (V_M, V_H) are overlapped with gene-regulatory datasets in the form of eQTLs (E_M, E_H; processed from GTEx, GeneNetwork, and specific mouse populations) and chromatin interaction studies (e.g., ChIA-PET experiments from ENCODE and gene-promoter interactions from the Eukaryotic Promoter Database). Association of regulatory variants and gene-regulatory information allows for the identification of putative gene targets. These datasets are harmonized within-species for mice (V_M, E_M, G_M, R_M) and humans (V_H, E_H, G_H, then related across species through orthologous gene targets (O_M, O_H) derived from homology resources like the Alliance for Genome Resources.

Fig 2. Multi-species genomic and epigenomic analysis. Species-specific gene, gene-regulatory, and variant-level data is harmonized from public resources. Using variant and gene annotations as input from post-GWAS annotation tools (e.g., FUMA, MAGMA, etc.), gene-regulatory components can be related across species via epigenomic modeling. Gene targets identified from epigenomic modeling can be used for further post-GWAS analysis with tools such as Enrichr, GeneWeaver, KnowEng, etc. Such analyses have numerous biomedical applications, such as the discovery of disease-relevant model organisms and traits.

Human Gene	Model Organism	Year of Publication	Model Organism Publication	Trait	Date	Human Publication	Trait
MPDZ	Mouse	2002	11978849	alcohol withdrawal	2009	19175764	alcoholism
MC1R	Mouse	2003	12663858	analgesia	2003	12663858	analgesia
OPRM1	Mouse/Rat	1994/1998	7982048/9512064	pain genetic variation/alcohol intake	1998	9756053/9689128	alcohol dependence/opioid binding, addiction
GAD1	Mouse	1994	8974318	alcohol withdrawal	2007	17034009	alcoholism
CHRM5	Mouse	2002	11900778	increased drinking	2004	15292665	schizophrenia
GABRB2	Mouse	2003	12490572	action of alcohol	1999	10195814	alcohol dependence
ALDH2	Rat	1991	2053491	alcohol drinking behavior	1982	7180842	alcohol metabolism Caucasian/Asian
ALDH1	Mouse	1996	4015840	alcohol metabolism inbred	1983	6354999	alcohol metabolism Caucasian/Asian
FAM53b	Mouse	2016	26581503	cocaine	2014	23958962	cocaine dependence
PPP1R1B	Mouse	1998	9694658	drugs of abuse	2006	16237383	amphetamine experience
CSNK1e	Mouse	1999/2005	10591541/16104378	amphetamine/cocaine induced stimulation	2006	16237383	amphetamine experience
COMT	Mouse	1975/1998	1185192/9707588	differential seizure susceptibility/KO social behavior	2003	12716966	methamphetamine brain response variation
DBH	Mouse	1991/1999 / 2000	1684202/10594079/ 10777779	altered norepinephrine and serotonin/seizure/alcohol	2000	10673769	cocaine-induced paranoia
DRD1	Mouse	1994	8001143	cocaine behavior	1997	9154217	addictive behavior
DRD4	Mouse	1997	9323127	supersensitive cocaine	1993	8216280/8268330	alcoholism/delusional behavior
DBH	Mouse	1992/2000	1542654 /11093800	ethanol induced	2000	10975602	smoking cessation
DDC	Mouse/Fly	1986/2006	3703899 /16783013	drug studies locomotor behavior	2005	15879433	nicotine dependence
HTR3A	Mouse	2001	11685380	conditioned place preference	2001	11207027	schizophrenia and bipolar
HTR5A	Mouse	1999	10197537	activity/lsd	2009	19328558	bipolar

ARRB 2	Mouse	1999	10617462	morphine analgesia	2006	16894395	ADHD
GRIN 3A	Mouse	2005	15866554	PPI	2009/ 2011	20016182/ 20084518	Alzheimer/nicotine
NRXN 1	Mouse	2009	19822762	PPI, learning, grooming	2005	16451640	COGA
HP1B P3	Mouse	2016	27460150	cognitive aging			
DAT1	Mouse	1998	10195128	cocaine IVSA	2001	11449401	ADHD
AP3M 2	Mouse	2014	24923803	alcohol preference/withdrawal			
HNRN PH1	Mouse	2015	26658939	methamphetamine			
GRIN 2B	Mouse	1996	8789948	abnormal startle	2000	10945659	ADHD, ODD and conduct disorder.
CHRN A3	Mouse	1999	10318955	megacystis-microcolon-intestinal hypoperistalsis	1998	9758605	epilepsy
CHRN B4	Mouse	2004	14996991	seizure	1998	9758605	epilepsy
CHRN A6	Mouse	2002	11927835	nicotine	2002	12195439	epilepsy
CHR M1	Mouse	2001	11752469	hyperactivity	2003	14504414	psychiatric symptomology
CHR M2	Mouse	1999	9990086	impaired drug response	2002	12116189	depression
CYP2 A6	Mouse	1989	2733794	altered metabolism	1998	9655391	nicotine metabolism
CYP2 B6	Mouse	2010	19923441	nicotine pharmacokinetics	1992	1736885	drug metabolism
NTRK 2	Mouse	1993	8402890	neonatal death	2005	15838534	eating disorder
MAP3 K4					2010	20624154	smoking
SHC3	Mouse	2005	15716419	spatial memory	2007	17179996	nicotine
DNM1	Mouse	2007	17463283	abnormal motor capabilities/ coordination / movement	2008	18806795	exercise-induced collapse
TAS2 R38	Mouse	2014	mousephenotypes.org	limb grasping	2005	15466815	taste
APBB 1	Mouse	2004	14689444	abnormal spatial learning	1998	10079843	Alzheimer disease

NRG3	Mouse	2016	27606322	abnormal behavior	2008	18708184	schizophrenia
DRD2	Mouse	1995	7566118	impaired coordination	1991	1832466	neuropsychiatric disorders

Tool Name	Description	Strategy
AnnoPred	Estimates PRS using genomewide variants that are differentially weighted based on the integration of evidence across GWAS summary statistics and multiple annotation resources for different tissue types, genomic features, and the functional assessment of SNPs.	Bayesian framework integration
DIAMOnD	This tool identifies potential variant to gene associations based on module inclusion. Uses an algorithm for detecting disease modules based on network connectivity.	Algorithm for network module analysis
ENCODE Screen	Useful for discovering the potential regulatory role of genetic variants using cis-regulatory elements from ENCODE data in human and mouse.	Database
FOCUS	Used to determine gene–trait associations from transcriptome wide annotation studies using LD among SNPs and eQTL weights embedded in a probabilistic model.	Probabilistic Systems Framework
FUMA	Online tool to visualize and aggregate positional, eQTL and chromatin interaction maps to perform enrichment analysis of human GWAS data. Can be used to associate genetic variants to target genes based on eQTL and chromatin interaction studies.	Tools pipeline and visualization
GeneNetwork	Set of variant, expression and eQTL multi-species tissue specific data sets used to link genetic maps to disease and phenotypes of interest.	Database, Statistical and Probabilistic Tools
GeneWeaver	Multi-species data integration tools that allows users to identify putative genes of interest based on shared or unique genetic or variant data of interest. Tools available to map, manage and analyze large datasets.	Bi-partite, k-partite, Combinatorics, Network Analysis
H-MAGMA	A modified version of MAGMA that extends gene-to-variant mapping by including long-range loci interactions predicted by Hi-C.	Statistical Multiple Regression Models
Harmonizome	Online resource for data integration from existing genomic resources.	Association Matrix, Machine Learning
HumanBase	Online tools for tissue specific gene and network interactions.	Association Network, Machine Learning.
KnowEng	Integrative analysis following formatted pipelines for knowledge discovery.	Knowledge Network, Machine Learning
LDPred-funct	Used to derive polygenic scores using multiple genetic variants . LDpred-funct estimates polygenic effects by employing a model that accounts for LD and identify trait-specific priors that are based on posterior casual associations.	Probabilistic modeling
MAGMA	Software tool used to assign GWAS identified variants to genes, based on physical proximity, and perform joint and conditional association models that examine gene-, gene-set, and interaction effects.	Statistical Multiple Regression Models

modENCODE	Collaborative data set for genomic functional elements across several species, used to define genomic regions and variants of interest.	Database, ModMine Toolset
Monarch	Semantic integration of phenotypic disease associations to identify underlying genes.	Knowledge Graph
PAINTOR	Used to determine SNPs to be tested for phenotypes of interest. Predicts the impact of multiple casual variants on genomic annotations by incorporating summary associations statistics, functional annotations, and LD statistics.	Probabilistic Systems Framework
psychENCODE	Collaborative data set for genomic functional elements, used to define genomic regions and variants of interest in the brain.	Database, ModMine Toolset
S-PrediXcan	Used to predict gene associations to disease using gene expression levels to mediate summary GWAS and measured transcriptome studies without the need to use individual-level data.	
SMR	Identifies genes with expression levels and pleiotropic associations with diseases of interest via the integration of GWAS variants and expression data derived from eQTL studies.	Mendelian Randomized Analysis
TWAS	Identifies expression-trait associations by creating putative transcriptome-wide associations derived by integrating gene expression measurement with GWAS estimated associations.