# THE DEVELOPMENT OF COMPUTATIONAL APPROACHES FOR SYSTEMS GENETICS AND ALTERNATIVE POLYADENYLATION STUDIES AND THEIR APPLICATION IN STUDYING GENETIC PREDISPOSITION TO ALCOHOL RELATED PHENOTYPES

by

RYAN LUSK

B.S., University of Minnesota Duluth 2013

A thesis submitted to the

faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Pharmaceutical Sciences Program

2021

This thesis for the Doctor of Philosophy degree

by Ryan Lusk

has been approved for the

Pharmaceutical Sciences Program

by

Peter L. Anderson, Chair

Laura M. Saba, Advisor

Richard Radcliffe

Katerina Kechris

Boris Tabakoff

Date: 08/20/2021

Lusk, R (PhD, Pharmaceutical Sciences)

The development of computational approaches for systems genetics and alternative polyadenylation studies and their application in studying genetic predisposition to alcohol related phenotypes

Thesis directed by Associate Professor Laura M. Saba

## ABSTRACT

Alcohol use disorder and alcohol related phenotypes are genetically influenced, complex traits. Complex traits have proved challenging for connecting genotype to phenotype because they arise from a number of genetic factors capable of interacting with each other and the environment. As a result, susceptibility to alcohol-related traits and diseases likely involves a network of genes across several biological systems. Moreover, individual susceptibility to complex traits is mainly due to variation in gene regulation rather than protein-coding sequence. Our current research sought to improve our understanding of the genetic architecture of alcohol related traits by developing and applying computational approaches that consider networks of genes and gene isoforms on studies involving these traits using animal models. Specifically, we first demonstrated an unsupervised, statistically based method for identifying networks of genes associated with a complex trait and applied this method to two alcohol metabolism phenotypes: alcohol clearance and circulating acetate levels. This systems biology approach – which integrated genotype, expression, and phenotype data – identified a candidate gene network that included the alcohol dehydrogenase genes, which have a well-documented history of influencing alcohol metabolism phenotypes, thereby supporting our methodology. We were also able to hypothesize how these alcohol dehydrogenase genes function with other genes in the network in the absence of alcohol. We next developed a machine learning algorithm to identify

polyadenylation sites of transcripts in the expressed transcriptome to characterize this gene regulation phenomenon. This algorithm, aptardi (alternative polyadenylation transcriptome analysis from RNA-Seq data and DNA sequence information), leverages both DNA sequence and RNA sequencing, and we demonstrate improvements in identification over single omics methods. Using this algorithm, we examined how alternative polyadenylation impacts predisposition to voluntary alcohol consumption, again by applying a systems biology approach. Here we were able to recapitulate genes we previously determined to be associated with this phenotype; however, by including polyadenylation structures, we were able to identify the specific gene isoforms associated with the trait. These studies provided additional insight into the genetic etiology of alcohol related traits, and application of these methods to future studies could likewise improve our knowledge.

The form and content of this abstract are approved. I recommend its publication.

Approved: Laura M. Saba

For my parents, Daniel and Molly, and my siblings, Kevin and Kelly. Also for my undergraduate research advisors, Drs. Anne Hinderliter and Elizabeth Austin-Minor.

## ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Laura Saba. In terms of research, she is a brilliant scientist, and I am privileged to have had the opportunity to learn from her. Beyond science, and as anyone who has collaborated with Laura can attest, she is a joy to work with. Additionally, she proved a fantastic advisor with a deft ability to assess the situation and coax the best out of her advisees. Despite these qualities, Laura remains extremely humble individual. This rare blend of qualities makes Laura a true "unicorn" in science that embodies the very essence what it means to be scientist. I would also like to thank my dissertation committee for all their guidance as I pursued my graduate education. Namely, thank you to Dr. Boris Tabakoff for lending your expertise on alcohol and the genetics of alcohol and acting as a "grandmentor" to me during my PhD. I appreciate our thoughtful conversations and pushing me to critically think in all aspects of my work. Thank you to Dr. Katerina Kechris for your assistance with the computational components of my research and guidance throughout my tenure. Thank you Dr. Richard Radcliffe for providing your expertise in genetics and also for the thoughtful and engaging questions posed to me during my committee meetings. And finally thank you to Dr. Peter Anderson. I appreciate you serving a chair of my committee and organizing the meetings. On a personal note, I am forever thankful that you took the time to talk with me about the challenges of graduate school, and about life in general. I could not have made it through without you. I would also like to thank Dr. Farnoush Banaei-Kashani and Evan Stene for serving as the machine learning experts on this project. Our meetings and conversations were essential to the successful completion of this work, and I very much enjoyed working with Evan in their lab at the downtown campus.

## TABLE OF CONTENTS

**Chapter**

# CHAPTER I

# REVIEW OF THE LITERATURE AND STATEMENT OF PURPOSE

This chapter will provide the reader the background information for the subsequent research chapters (Chapter II-IV). Briefly, this introduction will describe, in order, alcohol use disorder (AUD), the genetics of alcohol related phenotypes, animal models for studying these phenotypes, systems genetics approaches for interrogating the genetics of these phenotypes, the potential genetic role alternative polyadenylation (APA) on these phenotypes, and finally machine learning techniques to improve identification of APA transcripts in the expressed transcriptome.

## Alcohol use disorder

In the Americas, over half of the adult population consumes alcohol [1]. In 2016, the harmful use of alcohol resulted in some three million deaths worldwide, and mortality resulting from alcohol consumption is greater than that caused by diseases such as tuberculous, HIV/AIDS, and diabetes [1]. Alcohol also caused 7.2% of all premature deaths (persons aged 69 or younger) in 2016.

Excessive alcohol consumption can lead to AUD, which is defined as a problematic pattern of alcohol use accompanied by clinically significant impairment of distress [3]. Specifically, the Diagnostic and Statistical Manual for Mental Disorders, 5th edition (DSM-5) defines AUD as a pattern of alcohol consumption that leads to problems associated with two or more of the 11 potential symptoms of AUD within a 12 month period [3] (see [3] for symptoms). Approximately one third of American adults will meet the criteria for a diagnosis of AUD at some point in their lives [4].

## The genetics of alcohol related phenotypes

### Genetics of alcohol use disorder

The concept of AUD being at least partially under genetic control – resulting from the observation that the trait is characteristic of families – dates back to the nineteenth century.[5] Since 1960, scientifically designed studies utilized family [6], twin [7-13], and adoption [14-18] methods to address the heritability of AUD. Quantitative [19] and qualitative [20-22] analysis of these family, twin and adoption studies have since concluded that – despite major differences in methods – the results consistently demonstrate that genetics contribute approximately 50% to the development of AUD [19]. AUD is a complex disease where variations in large numbers of genes influence its risk and individuals can vary greatly in clinical symptoms that define their disease. Despite our longstanding knowledge of the underpinnings of AUD, much of the genetic component of AUD remains unexplained: the so-called "missing heritability."

### Endophenotypes for studying alcohol use disorder

One of the difficulties in identifying the genes responsible for AUD is clinical heterogeneity, which confounds results from genetic studies [23-27]. To overcome this, endophenotypes can be used. Endophenotypes, such as alcohol metabolism phenotypes and alcohol consumption, deconstruct current diagnostic categories for AUD into quantitative underlying traits that are more amenable to genetic studies [28]. The genetic components underlying endophenotypes are more easily identifiable since they will have a larger effect on the endophenotype than on AUD as a whole where diverse clinical symptoms and, most likely, biological processes constitute AUD.

**Alcohol metabolism**

Although several genes have been proposed to influence AUD, reproducibility has cast doubt upon nearly all of these genes [29-31]. The exceptions to this reproducibility problem are the alcohol metabolizing genes, namely alcohol dehydrogenase (*ADH*) and aldehyde dehydrogenase (*ALDH*). They have survived from the candidate gene era to today's era of genome-wide association studies (GWAS) [32-41], leaving little doubt regarding their genetic contribution to AUD.

The majority (95-98%) of imbibed alcohol (ethanol) is eliminated via metabolism to carbon dioxide and water.[42, 43] Although multiple metabolic pathways exist [44], hepatic oxidation in which alcohol dehydrogenase (ADH) converts ethanol to acetaldehyde, followed by rapid conversion of acetaldehyde to acetate by aldehyde dehydrogenase (ALDH) is the major metabolic pathway [43-46]. The genes linked to AUD encode the enzymes of this dominant metabolic pathway, i.e., ADH and ALDH, and therefore only the genetics of this dominant metabolic pathway will be discussed.

There are six *ADH* genes expressed in human that constitute the ADH enzyme family (a seventh has yet to be found as a protein *in vivo*) [32, 47]. These enzymes, which function as dimers [48-50], are further categorized into five distinct classes [44]. Of these, class I and class II are significantly involved in liver metabolism of ethanol [49, 51] – with the lower $K_m$ (~0.05-5 mM in human) [52], three-enzyme class I family being primarily responsible for metabolism of circulating levels of ethanol usually reached in ethanol imbibing human [50]. These proteins – ADH1A, ADH1B, and ADH1C – have over 90% amino acid sequence homology among the three[32] and, unlike the rest of the enzymes of the ADH family that only homodimerize, are capable of forming heterodimers with one another [32, 46, 50, 51, 53]. Based on mRNA levels,

ADH1B is the highest expressed subunit in human liver, with the other two being expressed about one third its levels [32]. As alcohol concentrations rise, the higher $K_m$ (~34 mM in human) [52] class II ADH4 enzyme starts to contribute significantly to ethanol metabolism [46, 50, 53, 54], albeit likely still to a lesser degree than the class I proteins (estimates of up to 1/3 that of the class I enzymes have been reported) [32]. At 22 mM alcohol concentration (0.1%; 0.08% is generally defined as legally intoxicated), the total ethanol oxidizing capacity in the liver is approximately 70% class I ADH1 enzymes and 30% class II ADH4 enzyme in humans [55]. The class II ADH4 enzyme RNA expression level is approximately two thirds that of ADH1B [32].

The ALDH superfamily consists of 19 known enzymes in human to date [56], but only three – ALDH1A1, ALDH1B1, and ALDH2 – are relevant to the oxidation of acetaldehyde produced from the metabolism of alcohol by ADH in the liver [57, 58]. Although its name suggests otherwise, ALDH2 is also a member of the ALDH1 subfamily along with ALDH1A1 and ALDH1B1; its longstanding name is due to its association with alcohol metabolism that has been grandfathered into the ALDH nomenclature [59]. This subfamily is characterized by their ability to synthesize retinoic acid from retinaldehyde and plays an important role in regulating retinoic acid signaling [59]. They all share at least 68% amino acid sequence identity and function as homotetramers [53]; however, their cellular locations [53] and substrate specificities [32] differ. ALDH1A1 is localized to the cytosol of the liver, whereas ALDH1B1 and ALDH2, which are 75% identical [60], reside in liver mitochondria and convert acetaldehyde to acetate in the mitochondrial matrix [53]. ALDH2 – among the top 100 genes expressed in liver – displays considerably higher expression in the liver than ALDH1A1 and ALDH1B1 [32] and, with respect to acetaldehyde, has the highest affinity, i.e. lowest $K_m$ (~0.2 mM in human) [61]. ALDH2 is responsible for the majority of acetaldehyde metabolism [61]. The cytosolic

ALDH1A1 has a much lower affinity for acetaldehyde, evidenced by its $K_m$ value being 900-fold greater in human than that of ALDH2 (~180 mM) [61], and is expressed at lower levels in the liver compared to ALDH2 [32]. Therefore, its contribution to acetaldehyde metabolism is likely only relevant when ALDH2 is inactive or saturated with substrate [32]. Similarly, while the ~55 mM (in human) $K_m$ value of ALDH1B1 [60] is lower than that of ALDH1A1, it exhibits the lowest expression in liver of the three enzymes and likely has a smaller role in acetaldehyde metabolism compared to ALDH2 [32]. Regardless, Singh et al. [62] demonstrated that *Aldh1b1* knockout mice have significantly higher blood acetaldehyde levels after intraperitoneal injection of alcohol, albeit at the relatively high dose of 5 g/kg, suggesting ALDH1B1 may contribute a measurable amount to acetaldehyde removal by the liver under some drinking conditions.

*Genetic studies on alcohol metabolism genes*

The genes encoding the ADH family are arranged head-to-tail in a 370 kb region on chromosome 4 in humans [53]. One of the first linkage studies for AUD from the Collaborative Study on the Genetics of Alcoholism (COGA) reported a broad risk locus encompassing this region.[63] A second sib-pair study in 1998 from a Southwest American Indian tribe provided further evidence for association on chromosome 4 [64]. Since then, several other studies have replicated the association between the region on chromosome 4 containing the genes encoding the ADH enzymes and AUD (or an alcohol-related trait) [65-70]. Some of these studies also identified a region on chromosome 12 where the *ALDH2* gene resides [69, 70].

Informed by previous linkage studies identifying the broad region on chromosome 4 containing the *ADH* genes, several follow-up positional candidate gene association studies on the *ADH* genes have been performed. Additionally, the genes encoding other products with obvious involvement in the actions of alcohol have been tested in several functional candidate gene

association studies. Of these, the strongest associations with AUD have been detected for markers across the *ADH* gene cluster and at the *ALDH2* gene [25, 33, 34, 45, 50, 51, 71, 72], and this remained true even when single studies were combined in meta-analyses to increase statistical power [73-75].

As with candidate gene studies, the most robust associations with AUD or alcohol-related traits from GWAS comes from the alcohol metabolism genes. Park et al. [37] and Quillen et al. [76] found significant associations with variations in *ALDH2*, and Frank et al. [41] and Gelernter et al. [35] likewise found significant associations with the *ADH* genes.

Unlike ALDH2, the overall evidence for association of AUD with *ALDH1A1* and *ALDH1B1* is weak. For instance, some low-frequency *ALDH1A1* variants have been nominally associated with alcohol-related traits [69, 77-79], but none have shown up in GWAS [32]. For *ALDH1B1*, associations have been found in some statistically underpowered studies [80-82], but these weak signals were not replicated in larger studies [83].

*Influence of genetics on alcohol metabolism gene function*

Multiple functional studies have explored how single nucleotide polymorphisms (SNPs) in the *ADH* genes that alter the amino acid sequences of their protein products, i.e., coding SNPs, related to their ability to oxidize ethanol. Most studies have focused on the class I ADH enzymes, likely because they are primarily responsible for alcohol metabolism in the liver. Similarly, the vast majority of researchers have been interested in protein variants that have clear functional consequences. To that end, multiple alleles of the class I *ADH* genes have been well studied with respect to their impact on alcohol metabolism: ADH1B*1 (the reference allele), ADH1B*2, and ADH1B*3. The ADH1B*2 variant has an arginine to histidine substitution at the 48$^{th}$ position of the amino acid sequence, and the ADH1B*3 variant has an arginine to cysteine

substitution at the 370$^{th}$ position [24]. Both substitutions are at an amino acid contacting the coenzyme nicotinamide adenine dinucleotide ($NAD^+$) that cause it to be released more rapidly during ethanol oxidation compared to the reference protein, which results in a 70- to 80-fold greater enzymatic turnover rate and increased rate of ethanol conversion to acetaldehyde [50]. Like *ADH1B, ADH1C* displays three alleles, of which the resulting protein of two have been extensively studied [50]. ADH1C*1, the reference, possesses an arginine at position 272 and isoleucine at position 350, while ADH1C*2 has a glutamine and a valine at these positions, respectively [50]. ADH1C*1 is about 1.5- to 2-fold more active than ADH1C*2 [84], again leading to increased acetaldehyde accumulation. The third allele, ADH1C*Thr352, encodes a protein with threonine at position 352 and is common in Native American populations [85].

The class II *ADH* gene has fewer identified coding SNPs compared to the class I *ADH* genes [32]. Stromberg et al. found a SNP that gives an isoleucine to valine substitution at 308 resulting in a less stable protein product; however, its $K_m$ value for alcohol was increased only slightly [86].

The reference ALDH2 enzyme possesses a glutamate at amino acid position 487 of the mature protein and is encoded by the *ALDH2*1* allele [32]. There is one significant genetic polymorphism of *ALDH2*, the coding variant *ALDH2*2*, with a lysine at this position instead [87]. Studies have demonstrated that even a single *ALDH2*2* encoded subunit in the ALDH2 tetramer renders the enzyme virtually inactive [88, 89]. Consequently, heterozygous individuals (that is, those with one *ALDH2*1* and one *ALDH2*2* allele) have little detectable activity for converting acetaldehyde to acetate and homozygous individuals have no detectable activity [88]. Moreover, the mutant tetramer is more rapidly degraded [90].

Many polymorphisms have been found for the *ALDH1A1* gene [53]. Those associated with alcohol-related traits lie in its promoter region [32], suggesting they may modulate *ALDH1A1* gene expression. In addition, functional polymorphisms of *ALDH1B1* have been described [91, 92] and at least one has shown associations with alcohol drinking habits and sensitivity [80, 81].

*Noncoding genetic variants in the alcohol metabolism genes and their association with alcohol phenotypes*

Many more noncoding, regulatory SNPs have been identified for the class I and class II *ADH* genes. Molecular studies of some of these noncoding SNPs have demonstrated that they affect expression of *ADH1* [93, 94] and *ADH4* [95], and may also affect the rate of ethanol conversion to acetaldehyde. Moreover, while coding variants in the *ADH* gene region and their kinetic properties have thus far been highlighted with respect to their impact on AUD, noncoding SNPs have been associated with risk for AUD as well. This aligns with the notion that the genetic architecture underlying complex diseases and traits is largely driven by gene regulation processes.[96-99]. For example, noncoding variations in and around *ADH4* are among the mostly widely replicated associations with AUD in several populations [53]. These include European-Americans [100], Brazilians [45], and another sample of European and African Americans [55]. Changes in *ADH4* gene expression may be responsible for the observed associations; noncoding SNPs in the promoter region of *ADH4* have been shown to alter its expression level [95, 101], and some of its strongest associations reside in the region near its 3' end [55]. Changes in the length of the 3' untranslated region (3' UTR) can have a significant impact on gene expression [102-104]. In addition, noncoding SNPs in the region of *ADH1A*, *ADH1B*, and *ADH1C* have been linked to AUD and drinking phenotypes [53], and noncoding SNPs that affect their gene

expression have been identified [94, 105, 106]. Many of these noncoding SNPs are in strong linkage disequilibrium with the coding variants associated with AUD delineated above, making it difficult to distinguish exact contributions [50]. Biochemical evidence suggests a role for at least the *ADH1B* coding variant, but this allele may also act by altering expression of one of the *ADH* genes [32], and still other coding variants found to be associated with AUD may in fact be linked because of the noncoding SNPs inherited together that alter expression.

**Alcohol consumption**

Despite the clear genetic connection between the alcohol metabolizing genes and AUD, much of the genetic influence on AUD remains undefined. To better understand the genetic architecture of AUD, some researchers have turned to endophenotypes such as alcohol consumption. While a diagnosis of AUD does not depend on the amount of alcohol consumed, alcohol consumption is considered an etiologic essential for the development of AUD [107, 108] and other alcohol-related problems (e.g., cirrhosis, pancreatitis, and upper gastrointestinal tract cancers [23]). Moreover, alcohol consumption is detrimental independent of AUD; for instance, alcohol use is the leading cause of preventable death in the United States [109], and many of the negative consequences of alcohol listed at the beginning of this introduction are due to excessive alcohol consumption and not AUD per se. Beyond its risk factor for disease and health impacts, costs associated with losses in workplace productivity, health care expenditures and criminal justice due to excessive alcohol consumption were $249 billion in 2010 [2]. Alcohol consumption is a genetically influenced [110, 111] complex trait [112, 113] lacking complete genetic characterization. Complex traits have proved to be a challenge for connecting genotype to phenotype because they arise from a number of genetic factors capable of interacting with each other and the environment [114]. Identifying the missing heritability of alcohol

consumption and AUD may lead to new treatment and prevention strategies since numerous studies have illustrated that the efficacy of alcohol pharmacotherapies depend on one's genetic makeup [115] and a substantial proportion of individual differences in disease susceptibility is due to genetic factors [116]. Four medications have been approved by The US Food and Drug Administration to treat AUD – disulfiram, oral naltrexone, extended release injectable naltrexone, and acamprosate [117]. Disulfiram acts by inhibiting ALDH, which produces the unpleasant effects of alcohol analogous to individuals with the ALDH2*2 polymorphism [118]. The naltrexone based therapies operate as opioid antagonists to reduce alcohol consumption [119]. Acamprosate has inhibitory effects at the metabotropic glutamate receptor 5 [120] and can reduce elevated glutamate levels in those with AUD, but its exact mechanism of action is not understood [118]. However, these therapies are only effective for a small proportion of individuals with AUD.

## Animal models for studying endophenotypes

Besides clinical heterogeneity, other challenges involved in identifying the genes contributing to AUD and its related endophenotypes include reduced penetrance, epistatic effects, contributions from gene-environment interactions [23, 24, 27, 50, 55, 121, 122], and the polygenic nature of AUD and alcohol consumption that necessitates larger sample sizes for genome-wide association studies (GWAS) [29, 123]. To circumvent some of these challenges, mice and rats have been used for decades to study many traits associated with AUD such as alcohol preference, alcohol sensitivity, and withdrawal sensitivity [124, 125]. The use of model organisms, such as inbred panels, has several advantages over human studies for analyzing the genetics of complex traits. For instance, inbred panels represent a permeant resource for trait mapping and analysis and provide a means to accumulate phenotype and genetic information

over time [126]. This likewise makes it possible to analyze biological replicates and perform validation studies. Additionally, genotyping of each strain is only required once [127], multiple individuals from each strain can be phenotyped to reduce the impact of technical and environmental factors on the phenotype(s) [128], and multiple phenotypes can be assessed on the same genetic background. Environmental factors can also be controlled in organisms to eliminate their impact or even systematically manipulated and/or statistically accounted for to examine gene-environment interactions [28]. Additionally, reducing the contribution of environmental variation to the total error variation can increase statistical power [28]. Inbred animals are also homozygous at all SNPs, providing additional statistical power and simplifying interpretation. Animal models also allow for utilizing both forward and reverse genetics approaches in complex traits and, in general, provide greater control over experimental design.

**The HXB/BXH recombinant inbred rat panel and its utility for genetic studies**

One such inbred panel, the HXB/BXH recombinant inbred (RI) panel, has been used to study several alcohol related phenotypes [129-133]. The HXB/BXH RI rat strains were derived by gender-reciprocal crossing of a spontaneously hypertensive rat (SHR/Ola) with the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub) [134]. Specifically, these progenitor strains were mated to produce F2 hybrids, each of which F2 animal contains a practically irreproducible unique combinations of genes due to both independent segregation of maternal and paternal chromosomes and recombination between homologous chromosomes during meiosis. Subsequent inbreeding of randomly chosen pairs of F2 animals and brother-sister mating for at least 20 generations yield individual, isogenic strains [134]. Each strain carries unique paternal-maternal gene combinations, akin to the F2 animal and, since gender-reciprocal crossing was utilized – two different sets of strains (denoted as HXB or BXH) that differ in the

source of mitochondrial DNA and the Y chromosome [134]. The HXB/BXH RI panel consists of 30 lines that have been inbred for >80 generations [135]. This panel has been used to investigate a wide variety of complex traits [136] and many complex traits, such as the alcohol related phenotypes studies here, have been found to vary across this panel and are thus amenable to genetic studies [132, 137-142].

*Comparison of human and rat alcohol metabolizing genes*

Historically, the nomenclature used to describe alcohol metabolizing genes in different organisms was disjointed. Since in this present work we utilized the HXB/BXH RI panel to study alcohol metabolizing genes, we clarify the nomenclature used here and how they relate their human counterparts. Namely, we use the recommended nomenclature that has recently been proposed based primarily on amino acid sequence homology and secondarily upon catalytic properties or expression patterns to orthologs [143]. Furthermore, the protein and gene names that provide the same name for the same protein/gene in different species are used [144]. So, for example, rat *Adh1*/Adh1 and *Adh4*/Adh4 are analogous to the human *ADH1*/ADH1 and *ADH4*/ADH4 genes/enzymes. Rodents (i.e. rats) possess *Adh* genes for each of the human *ADH* genes; however, unlike humans, rodents express a single class I *Adh1* gene instead of three isoforms [144]. Likewise, the $K_m$ value of rat Adh4 may be somewhat greater than its human counterpart [145]. Also like humans, the rat *Adh* genes are located near one another and in the same transcriptional order, albeit on chromosome 2 rather than chromosome 4 [146].

## Systems genetics

### Benefit of systems genetics

Many of the human genetics studies of AUD mentioned previously were genome-wide association studies (GWAS). GWAS aim to associate genotypes of SNPs within a population to

the phenotype in the given population and are commonly employed for studying complex traits [147]. As a forward genetics approach, the methodology of GWAS is as follows: the complex trait is first measured in a population and then the genetic components influencing differences in the phenotype are assessed. In other words, the study begins with a particular biological phenomenon and asks which genes are necessary to observe the phenomenon [148, 149]. The major advantage here is its unbiased nature; no assumptions are made regarding the molecular basis of the phenotype in question [150], which enables identification of previously unknown genes associated with the trait. From a medical perspective, the ultimate goal of GWAS are to identify casual variants of a phenotype and elucidate the functional effects and the biological pathways they impact for drug design [151]. While many loci have been independently identified, understanding the molecular mechanism(s) from GWAS alone is difficult. For instance, associated SNPs are often part of a larger region of correlated variants, thereby making it difficult to identify the causal variant. Additionally, many associated SNPs lie in intergenic regions, suggesting the underlying biological mechanism is regulatory [151]. Finally, SNPs are each assess independently although genes may interact with one another to produce a phenotype or the perturbation of any one of a collection of genes that function in the same pathway may produce similar phenotypes.

Systems genetics seeks to characterize the connection between genotype and phenotype by integrating intermediate phenotypes [152]. For example, incorporating transcript expression data allows for characterizing their contribution to complex disease and can contribute to a better mechanistic understanding of trait variation [153]. Expression data can also be used to ascertain networks of coexpressed genes. The concept of evaluating genes in the context of networks aligns with the observation that biological functions arise from complex interactions; each gene

is estimated to be involved with four to eight other genes [154] and involved in 10 biological functions [155]. Therefore, determining networks of genes associated with a complex trait can improve our understanding of the biological processes involved and aid in identifying potential therapeutic targets [156]. Another important advantage of network methods is that they greatly alleviates the burden of multiple testing by focusing on modules rather than individual transcripts [157].

**Overview of weighted gene coexpression network analysis**

Weighted gene coexpression network analysis (WGCNA) [158] is a popular method for generating networks of coexpressed genes/transcripts, i.e., modules. The overall result of WGCNA is the grouping of genes into modules based on similarity of gene expression. This is done in three broad steps: 1) determining the coexpression strength between all gene pairs, 2) determining the proximity of genes, and 3) identifying mutually exclusive clusters of genes based on the proximity measure. To accomplish the first step, an adjacency matrix is calculated that indicates the network connection strength between all pairs of genes. The coexpression similarity is an intermediate quantity for determining connection strengths that is typically defined as the absolute value of a correlation coefficient, by default the Pearson product moment correlation coefficient. A soft thresholding procedure using the coexpression similarity values is then used to transform the coexpression matrix into an adjacency matrix to mimic a scale-free network. While hard thresholding can be used, i.e., genes with a coexpression similarity above a threshold are assigned one and those below a threshold are assigned zero, soft thresholding is a biologically motivated method [159] that also limits information loss [160]. In soft thresholding, coexpression similarity values are raised to a power, i.e., soft threshold, to mimic a scale-free network. Scale-free topologies imply that few genes are highly interconnected “hub” genes, i.e.,

these genes are coexpressed with many other genes, while many other genes have few connections [161]. To ensure connections between genes are robust, the adjacency matrix is transformed into a topological overlap matrix. While the adjacency matrix considers the coexpression similarity between pairs of genes in isolation, the topological overlap matrix considers pairs of genes in relation to other genes in the network [162]. Namely, genes have high topological overlap if they are connected to roughly the same group of genes in the network, i.e., share connections with third party genes. Finally, this topological overlap measure is used as input for hierarchical clustering to determine coexpressed modules. This method uses unsupervised clustering, and modules are defined as branches of the resulting clustering tree using a dynamic branch cutting approach.

Using WGCNA in combination with the multiple data sources available in systems genetics studies, i.e., expression data and genotype data, can also be used to reduce the number of false positive networks associated with a phenotype. For example, one approach to identify candidate networks (modules) is to require the satiation of three conditions: 1) the module eigengene had to be significantly correlated with the quantitative phenotype measured across strains, 2) the module eigengene had to have a statistically significant quantitative trait loci (QTL), and 3) the module eigengene QTL had to reside within the 95% Bayesian credible interval of a significant or suggestive physiologic/behavioral QTL.

## Beyond genes: alternative polyadenylation

### Review of polyadenylation and alternative polyadenylation

Thus far, the systems genetics approaches used for identifying gene networks associated with alcohol metabolism and alcohol consumption have only been explored at the gene level and not thoroughly evaluated in the context of differences in transcript structures such as different 3’

ends, i.e., APA. The mature 3' ends of nearly all eukaryotic mRNAs, are created by a two-step process: 1) endonucleolytic cleavage of the pre-mRNA and 2) synthesis of a polyadenylate tail onto the upstream cleavage product [104]. In other words, polyadenylation (polyA) defines the 3' ends of transcripts. The molecular machinery involved in polyA is extensive and is a continuous area of research that is beyond the scope of this dissertation.

Individual susceptibility to complex diseases is mainly due to variation in gene regulation rather than protein-coding sequence [96-99]. In recent years, it has become evident that alternative polyadenylation (APA) is an extensively used post-transcriptional mechanism [104]. For instance, an estimated 70% or more of human genes encode multiple transcripts derived from APA [163]. APA enables a single gene to encode multiple mRNA isoforms with alternate 3' ends through usage of different polyA sites and can be thought of as a subtype of alternative splicing. Changes in the length of the 3' UTR can have a significant impact on gene regulation via modified mRNA stability, localization, and protein translation efficiency [102, 164]. Yet the importance of APA was only recently realized and, by extension, our biomedical knowledge of APA as well [165-168]. Nonetheless, changes in expression of APA transcripts have already been associated with disease [169], and is increasingly being recognized as a risk factor in complex diseases [170].

The 3' UTR of mRNAs contain many binding sites for regulatory RNA-binding proteins (RBPs) and microRNAs (miRNAs) [164, 171, 172]. For example, perhaps the most studied consequence of APA is its effects on miRNA binding; in mammals, more than half of conserved miRNA target sites are located between APA sites [173, 174]. Interestingly, studies on alcohol indicate that "master regulator" miRNAs control the development of tolerance towards alcohol and influence other complex substance use disorders and that miRNAs may also mediate other

alcohol pathologies [175, 176]. Due to the highly interactive processes impacted by differential expression of APA transcripts, a systems genetics approach that enables network modeling [177] may aid in our understanding of the genetic underpinnings of alcohol phenotypes related to APA. Indeed, susceptibility to alcohol-related traits and diseases likely involves a network of genes across several biological systems [115]. However, a systems genetics study with APA isoforms necessitates their transcriptome-wide quantitation side-by-side with other RNA molecules.

**Current methods for identifying polyadenylation**

There are three broad sequencing technologies utilized to identify polyA sites: 1) polyA signals within DNA sequence, 2) changes in read coverage from short-read RNA sequencing (RNA-Seq), and 3) direct 3' end RNA sequencing, but each possesses inherent limitations for sample-specific identification of polyA sites.

*DNA sequence-based methods*

One category of algorithms have sought to capitalize on the wealth of research connecting specific strings of DNA nucleotides, or DNA sequence elements, to polyA (see Tian and Gaber [178] for a detailed review). Most of these methods, e.g. DeepPASTA [179], Omni-PolyA [180], and Conv-Net [181], deploy machine learning approaches but conspicuously do not consider *in vivo* expression. These methods identify polyA sites that have the potential to be expressed using DNA sequence but, since they do not incorporate any expression data, cannot distinguish whether the site is actually expressed in a sample. This is an important consideration because polyA is a dynamic gene regulation mechanism. In other words, expression of potential polyA sites and APA phenomena can differ based on a multitude of factors. This includes cell growth, cell cycle, differentiation, and development [182]. Additionally, pathological situations such as cancer, immune response, inflammation, and viral infection can modulate 3' end

processing [182]. Thus, *in vivo* expression of alternative polyadenylation isoforms varies based on a myriad of factors such as tissue, physiological, and disease states [104, 183] and, furthermore, displays tissue specificity [184, 185]. As a result, characterization of polyA sites in the transcriptome should consider sample specific expression.

*Short RNA sequencing based methods*

Next generation RNA-Seq has become the standard technology to profile the expressed transcriptome. The resulting short reads are used by transcriptome assemblers to produce a genome-scale, sample-specific transcriptome map. Transcriptome assembly has proven a powerful approach to assess the transcriptome, but accurate determination of polyA sites from short read RNA-Seq alone is a known shortcoming [104, 186-189]. Unlike splice junctions which can be precisely located via reads that span the junctions, polyA sites are characterized by a gradual drop off in coverage [190]. For assemblers that harness prior annotation to guide the reconstruction, often the annotated polyA site assumed by the assembler is not correct [191]. While many transcriptome assemblers have been developed – each with its own design – to our knowledge none have demonstrated competence at annotating 3' ends. Some assemblers, e.g., Cufflinks [171]/StringTie [191], construct a minimum path RNA-Seq cover to the position where there is zero read coverage to annotate the 3' end of a transcript [190, 192]; but, since reads can be derived from precursor mRNA [193], this often results in an overestimation of polyA sites. Others, e.g., Scripture [194], calculate scan statistics above genomic background to define transcript structures, but this approach tends to produce biased estimates of polyA sites and in general is not well-suited for defining 3' ends [195]. Importantly, these strategies are only capable of producing a single transcript stop site per intron chain structure, which tends to be the distal polyA site, thereby missing imbedded proximal polyA sites, i.e., APA isoforms [190]. The

challenge of accurately identifying polyA sites is apparent to both the developers and those evaluating assemblers by way of allowing for error at 3' end predictions when assessing accuracy [191, 196].

Acknowledging the challenges of annotating polyA sites and design shortcomings of transcriptome assemblers to do so, researchers have developed supplemental tools to characterize APA dynamics from RNA-Seq. Chen et al. [197] provided a comprehensive critical review of these methods which we will briefly highlight here. There are three main methods for characterizing polyA sites and/or quantifying APA dynamics. Those that require *a priori* annotated polyA sites, e.g., MISO [198], QAPA [199], and PARQ [200], cannot identify *de novo* polyA sites. Others such as Kleat [201] and ContextMap 2 [202] utilize reads with strings of adenosines not derived from a DNA template, i.e., polyA tails. However, studies have demonstrated that polyA reads are scarce in RNA-Seq data [203, 204], resulting in low sensitivity and missing weakly expressed polyA sites. Finally, those that consider fluctuations in read coverage near the 3' ends of transcripts, e.g., DaPars [205], APAtrap [206], and TAPAS [207], are largely interested in single gene APA switching and/or quantifying differential APA usage between two groups of samples rather than producing a complete transcriptome. Also – as noted by Chen et al. [197] – these tools are not user-friendly; specific input formats are required and outputs are not readily integrable into downstream studies.

*Targeted methods*

An alternative approach is to directly capture 3' ends of mRNA with sequencing technology e.g., PolyA-Seq [208], 3' READS [209], PAS-Seq [168], etc. (see Shi [163], Elkon et al. [210], and Ji et al. [211] for a complete review). These methods are accurate at characterizing the genomic locations of polyA sites; however, whereas RNA-Seq data are widely available, 3'

sequencing data represent only a small fraction of available sequencing data and is costly and labor intensive to produce [190, 197]. In PolyA-Seq, library construction consists of the following: 1) reverse-transcription primed with an oligonucleotide consisting of the following: a universal sequence for downstream polymerase chain reaction (PCR), a string of 10 thymidines followed by a base other than thymidine, then a random base (i.e., any of the four bases), 2) second strand synthesis using random hexamers linked to a second PCR anchor, and 3) nested PCR to add Illumina specific adaptors while preserving strand orientation. For sequencing a primer ending in 10 thymidines is used, which results in a read whose sequence starts with resulting the base immediately upstream the string of adenosines [208].

## Machine learning

### Overview of machine learning

Combining the information afforded by both RNA-Seq and DNA sequence in a multi-omics approach may improve upon current methods for identifying polyA sites. To combine these data and make predictions on the locations of expressed polyA sites, a machine learning approach will be used. Machine learning can be defined as the practices and set of tools to give the ability to computers to find patterns in data without being explicitly programmed [212]. In other words, in machine learning the computer is not told what to look for; instead, data are provided to machine learning algorithms that then find patterns and correlations in the data to draw future conclusions and probabilities (i.e., predictions) when given new data. Machine learning tends to focus on making predictions without necessarily understanding the underlying mechanisms. This contrasts inferential statistical models, which typically focus on inference by creating and fitting a project-specific probability model [213]. One tangible example is using linear regression in (supervised) machine learning vs inferential statistics. In machine learning,

best practice is to train the model on a subset of the data and then tested on a held out set of data. The overall goal of this is to evaluate how the model will perform when making predictions (by evaluating its performance on the test set). In other words, machine learning focuses on the prediction accuracy. For the inferential statistical models, the line that minimizes the mean squared error is determined. The point here is to characterize the relationship between each predictor and the response in context of given dataset rather than use the resulting model for predictions on future datasets. Regardless, the statistical model can still be used to make predictions. Indeed, the demarcation between machine learning and inferential statistics is blurry, leading some to use the term statistical learning.

Machine learning has been widely applied in genomics research, which consists of large datasets that are particularly suitable for this type of analysis. For instance, genomic sequence elements, i.e., DNA sequence, have become a popular choice for applying machine learning. Beyond recognizing sequences in DNA, machine learning has been applied to other high throughput sequencing assays including RNA-Seq data, DNase I hypersensitive site sequencing, and micrococcal nuclease digestion followed by sequencing [214].

The process of machine learning is dynamic and can be divided into somewhat arbitrary steps. For example, one may define the steps as data collection and preparation, choosing a model, training the model, evaluating the model, and making predictions, while another researcher may include formatting the data and feature engineering as additional steps. Of note, these steps are typically done cyclically; for example, evaluation of the model may result in choosing a new model for training.

**Supervised and unsupervised machine learning**

Machine learning can be divided into two major paradigms: supervised and unsupervised. Supervised machine learning uses a training set that consists of data that have been labeled, i.e. the model has example input:output pairs from which it can learn how to map the input to the output. In contrast, unsupervised machine learning does not use labeled data, i.e., it does not possess the output, or answers, to input data [215]. In a broad sense, the primary goal of supervised machine learning is to produce a predictive model, whereas unsupervised machine learning is used for finding patterns in data. Supervised machine learning can be used for both regression and classification tasks, which refers to whether the labeled data are continuous (regression) or categorical (classification). An example of a supervised machine learning model for regression is to predict housing prices based on factors such as lot size, location, etc. using known housing prices for training. A classification task may be to predict house color using the same features and known house colors. Unsupervised machine learning is most often applied when the goal is clustering or dimensionality reduction. An example of using unsupervised machine learning for clustering is to segment customers based on their purchases; there is no inherent clustering that customers belong to (i.e., output), but grouping similar customers may aid in targeting marketing or experiences to groups of customers that share common purchases and thus may likely enjoy similar marketing/experiences. Unsupervised machine learning for dimensionality reduction is often used to select/extract features for supervised prior to performing supervised machine learning. This is done to remove uninformative or correlated features and also to limit overfitting. Using the housing example once again, there may be many features available for predicting the house price, and unsupervised may be used to select features

or extract features that are then used in supervised machine learning rather than the entire feature set. The present work focuses on a supervised machine learning for classification.

**Artificial neural networks**

One class of machine learning models are artificial neural networks, or simply neural nets (Figure 1.1). These networks, loosely based on the brain, consists of artificial neurons (i.e., nodes) which are densely connected. The strength of the connection between two nodes (i.e., their proximity) is assigned a quantitative weight. Commonly, neural nets are connected via an input layer, one or more hidden layers, and an output layer. Input features are fed to the input layer. Hidden layer(s) perform nonlinear transformations on the data received from the input layer. Finally, the output layer produces predictions from the information received by the hidden layer(s).

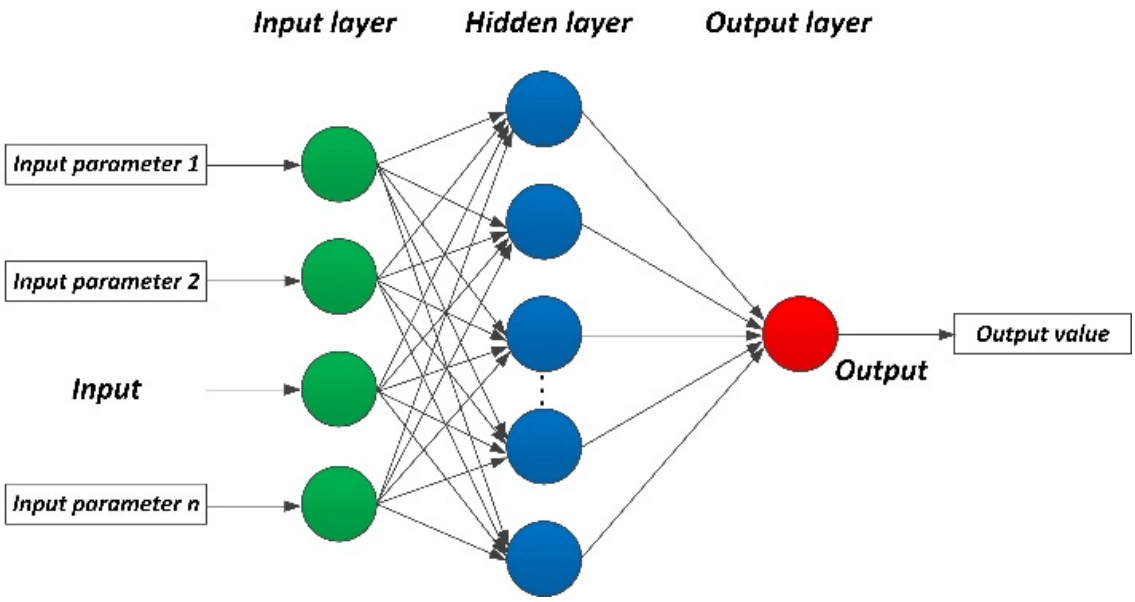



**Figure 1.1. From [216].** The structure of a neural network. © 2018 IEEE.

Interconnected nodes receive inputs from their connections, and each input has an associated weight. The weight is multiplied by the input for each connection, and this output for each connection is summed to give a single number for each node. This sum is passed through a

node's activation function to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome. During training, weights are randomly assigned a value. Training data are passed through the nodes and nonlinearly transformed through this activation process until reaching the output layer. The weights are then continually adjusted so that the output matches the labels, which is essentially the learning process.

There are three common types of supervised learning neural nets: feed forward nets, convolutional neural nets, and recurrent neural nets. Feedforward nets are often simply referred to as artificial neural nets and represent the core architecture of artificial neural nets. These networks only process information from input to output. Neurons in a layer do not connect to each other, and each neuron in one layer connects to all neurons in the next layer. These types of networks are appropriate for mapping input to output akin to other machine learning methods but benefit from the artificial neural network architecture in that they are able to abstract complex patterns through use of nonlinear transformations and additional numbers/layers of nodes.

Nodes in convolutional neural nets, on the other hand, are only connected to a subset of nodes in the previous layer [217], i.e., local connections. In other words, nodes in one layer only get information from the nodes sharing that region in the previous layer. This greatly reduces the number of parameters and thus the complexity of training. Additionally, the local connection weights are fixed for the entire nodes in the next layer [217], which provides the opportunity for convolutional neural net to detect and recognize input regardless of its spatial position in the data. These characteristics make convolutional neural nets ideal for image analysis.

Recurrent neural nets are an extension of feedforward neural nets that introduce the notion of time by the inclusion of edges that span adjacent time steps or sequence. (Figure 1.2A).

These edges, termed recurrent edges, form cycles, which include those of length one that represent a node connected to itself across time [218]. At a given time, nodes with recurrent edges receive information from the current data point and also from hidden node values from the previous state. In other words, the input of previous time steps can influence the output at the current time step. The time step relationship can be visualized by unfolding the network as depicted in Figure 1.2B. Instead of interpreting the recurrent edges as cyclic, this allows for visualization of the recurrent neural net as a deep network with one layer per time step and shared weights across time steps [218]. These networks are often employed on sequential data.

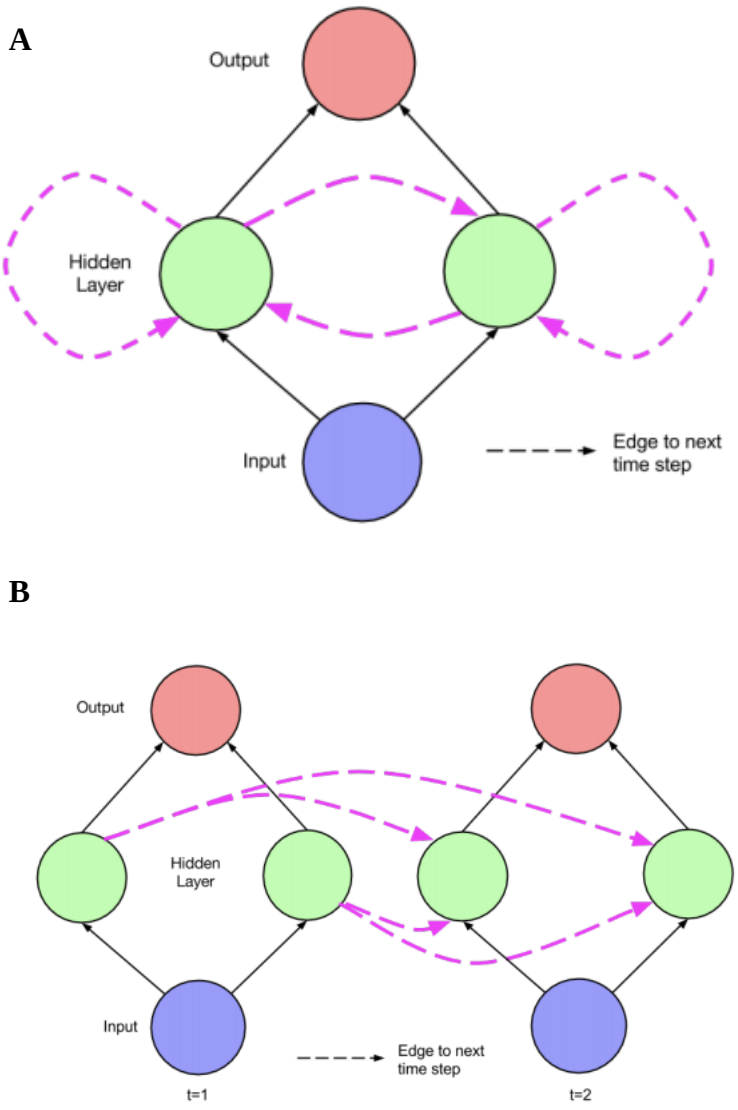



**Figure 1.2. From [218]. (A)** A simple recurrent network. At each time step t, activation is passed along solid edges as in a feedforward network. Dashed edges connect a source node at each time t to a target node at each following time t + 1. **(B)** The recurrent network unfolded across time points.

**Bidirectional long short-term memory networks**

Early recurrent neural nets, while effective at learning short term time dependencies, were ineffective at learning long term temporal dependencies due to the vanishing/exploding gradient problem [219]. During learning, backpropagation of the loss into the neural net is required so that nodes can adjust their weights to match the output during training.

Backpropagation is an efficient, dynamic programming method that uses the chain rule to work backward from the last layer (or timestep) to the first layer (or timestep) to calculate the error gradients with respect to each weight, i.e., partial derivatives. However, as this gradient is backpropagated through time using the chain rule, its magnitude is multiplied over and over again; as a result, there is exponential convergence to zero for small values (i.e., vanishing gradient) or accumulation of extremely large error gradients (i.e., exploding gradient). In either case, long term weights are not adjusted properly, and thus long-term dependencies are not accurately captured. To rectify this issue, the long short-term memory network (LSTM), a subclass of recurrent neural nets, was developed [220]. In this architecture, memory cells are introduced that replace the hidden nodes in traditional recurrent neural networks (see Appendix D for additional details). This type of architecture is useful for sequential data with a predefined endpoint, i.e., it accounts for both past and future sequence elements. An extension of LSTM, the bidirectional long short-term memory network (biLSTM) [221], has two layers of hidden nodes. The two layers are both connected to the input and output. These layers differ in that one has recurrent edges from past time steps, while the other has recurrent edges from future time steps.

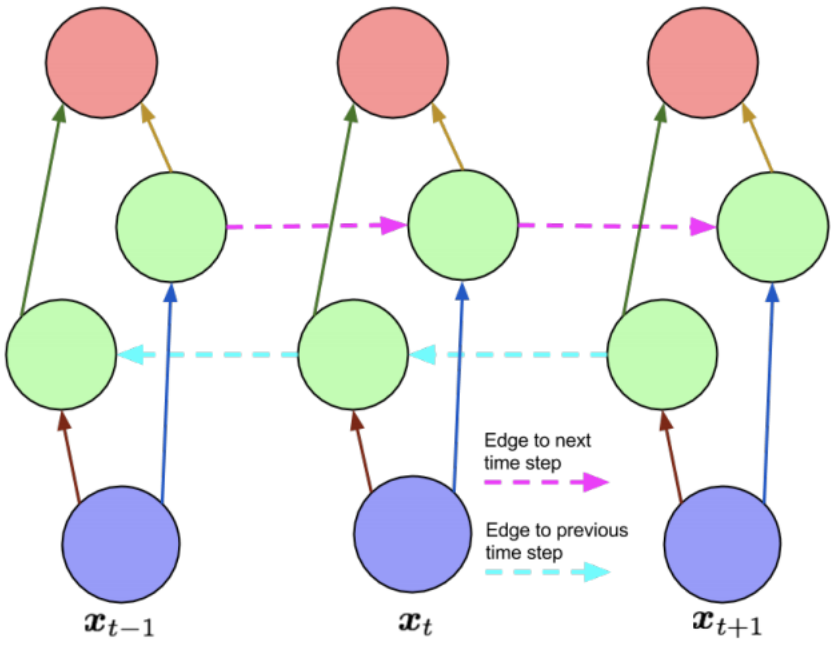



**Figure 1.3. From [218]. A bidirectional long short-term memory network unfolded in time.**

**Training machine learning models**

Two main aspects of the machine learning model that need to be defined for training are the loss function and optimization. The loss function defines what the model attempts to minimize during learning. For instance, a common choice for supervised machine learning when the outcome is continuous is the mean square error. This is calculated by taking the average of the square of the difference between the original and predicted value of the data. Similar to regression, model parameters are adjusted to minimize the mean squared error. In the case of binary classification, one method is binary cross entropy, or log loss (Eq. 1.1)

$$Log\ loss = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log \hat{y}_i + (1 - y_i) * \log(1 - \hat{y}_i) \quad (1.1)$$

where $N = sample\ size, y_i = acutal\ output\ at\ i, \hat{y}_i = predicted\ probability\ at\ i,$ Overall, binary cross entropy is the negative average of the log of corrected predicted probabilities.

Optimization defines how the model minimizes the loss function. A popular choice, especially for neural nets, is stochastic gradient descent (see Appendix D for more details).

Another important concept during training is the bias-variance tradeoff, which analyzes the expected generalizability of a machine learning model. Bias and variance are two contributing factors to the overall error. In particular, bias is the error of the target function (i.e., model parameters) from the real values, whereas variance is a measure of the deviation of the target function when using different training samples. In supervised machine learning, high bias is analogous to underfitting, meaning the model is unable to capture the underlying pattern of the data. In contrast, high variance equates to overfitting and indicates the model is capturing noise along with the underlying pattern of data. See Appendix D for additional discussion on underfitting and overfitting.

A common technique employed during the learning process is partitioning the data into train-validate-test splits. The test set is used to get an unbiased measure of model performance (i.e., its generalizability) using defined performance metrics, which will be discussed in the next section. Likewise, the validation set is used to evaluate model performance, but with the intent of tuning hyperparameters during the training process. Hyperparameters are parameters whose values are set by the user prior to the learning process, e.g., the number of nodes in an artificial neural net, the number of layers, the magnitude of the regularization parameter, the loss function, the optimization method. By training a machine learning model using different hyperparameters and then evaluating how these influence the performance on the validation set, one can get an unbiased measure of the best set of hyperparameters amongst those explored. However, partitioning a portion of the data solely for validation reduces the amount of data that can be used for training. Furthermore, only a single estimate of performance per set of hyperparameters can

be calculated. In response to this, many researchers split the data into only a training and testing split and perform cross validation using the training set. For example, in K-fold cross validation, the training set is split into k subsets, and the training-validation procedure is repeated k times. For each fold, a random k subset is used for validation, while the remaining k-1 subsets are used for training. Often times leave one out cross validation is employed, which is the logical extreme of K-fold cross validation in which a single data point is held out for validation and the number of training-validation procedures equals the number of data points. These cross-validation methods still enable optimization of hyperparameters while also allowing for average performances to be calculated for a given set of hyperparameters and more data to be used during training.

**Performance metrics**

Since supervised machine learning relies on labeled data, i.e., data where the outcome is known, performance can be directly assessed by comparing the predicted to expected output (i.e., label). Performance metrics are used during training to optimize the model and also to evaluate the generalizability of the model on unseen data. Many performance metrics are available, and the choice can depend on the task, the data, and the intended use. Here we will focus on performance metrics for binary classification with a particular emphasis on unbalanced data.

There are four possible outcomes when comparing predictions to actual values in binary classification, which can be summarized using a confusion matrix (Figure 1.3). Namely, a prediction of yes on an actual value of yes is termed a true positive (TP), a prediction of no on an actual value of yes is a false negative (FN), a prediction of yes on an actual value of no is a false positive (FP), and a prediction of no on an actual value of no is a true negative (TN).

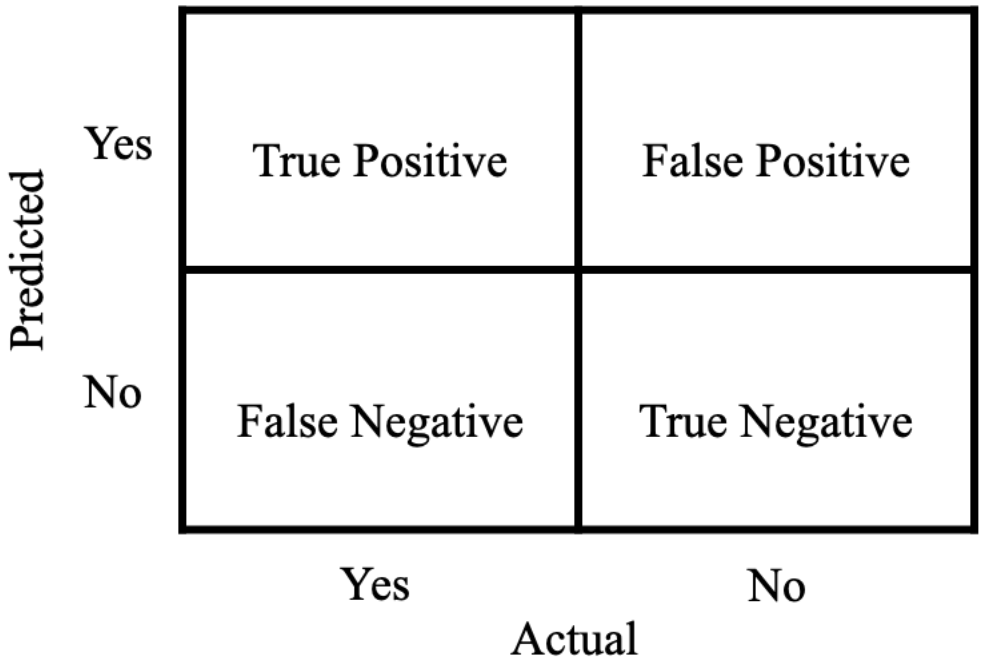



**Figure 1.3. A confusion matrix representing the four possible outcomes when mapping predictions to output in binary classification.**

The number of TP, FP, TN, and FN depend on the specified probability threshold that dictates whether the output probability (i.e., prediction) is classified as yes or no. In the simplest case, the accuracy can be calculated at a given probability threshold by summing TP and TN and dividing this by the total number of predictions. However, accuracy can be a poor measure of model performance. For example, imagine a dataset with 100 data instances where the number of yes values is 2 and the number of no values is 98. A null machine learning model that simply predicts no in all cases would have 98% accuracy but would not be able to identify any yes values, which may be of interest.

Inevitably, improvement of one parameter in the confusion matrix often comes at the cost of another. For instance, increasing the number of true positives can be achieved by decreasing the probability threshold for labeling a prediction yes; however, this will likely also increase the number of yes predictions that are actually no, i.e., the false positive rate. To visualize these

tradeoffs, several metrics can be derived from TP, FP, TN, and FN and plotted against one another as a function of probability thresholds including sensitivity/recall, specificity, and precision. Sensitivity (also referred to as recall and true positive rate) measures the total number of yes values are captured by the model (Eq. 1.2). In other words, a sensitivity of one indicates all yes values were identified by the model. Specificity (also referred to as true negative rate) does the same but for the no values (Eq. 1.3).

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{1.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{1.3}$$

In a receiver operating characteristic (ROC) curve, the false positive rate, or one minus the specificity (Eq. 1.4) is plotted against the sensitivity as a function of the probability threshold.

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{FP + TN} \tag{1.4}$$

The area under the curve is used to generate a single number to evaluate the model. However, the ROC curve is uninformative in the case imbalanced classes, i.e., many more no values than yes values. In this case, the false positive rate tends to stay at comparable, small values because the number of TN dominates its value. Instead, the precision-recall curve is often used when dealing with class imbalance. Here precision (Eq. 1.5) is plotted against recall (i.e., sensitivity) as a function of the probability threshold.

$$Precision = \frac{TP}{TP + FP} \tag{1.5}$$

Precision indicates out of all ‘yes’ predictions, how many are actually yes. Unlike the ROC curve, which incorporates the number of negatives via TN in the false positive rate equation, the precision-recall curve is concerned only with the yes class. Like the ROC curve, the area under

the precision-recall curve, or average precision, can be used to summarize performance in a single number.

While the precision-curve indicates model performance over a range of probability thresholds and is useful for the imbalanced case, one may be interested in its performance on imbalanced data at a single threshold. For this, the F measure can be used (Eq. 1.6).

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall} \tag{1.6}$$

The β term allows for weighting precision and recall differently; for example, setting it equal to two weights recall twice as importantly as precision. In the case of weighting the precision and recall equally, referred to the harmonic mean between precision and recall, the β term is set to one and Eq. 1.5 reduces to Eq. 1.7.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{1.7}$$

As we showed above, the choice of performance metrics may depend on the data (balanced vs imbalanced). Likewise, the intended use of the machine learning model may also dictate which metric is most important. For instance, in the case of cancer screening, it may be more important to have a high recall to ensure all instances of cancer are identified for a more invasive follow up examination, even if this means also including some patients without cancer in the follow up. In contrast, determining which websites are safe to visit may be more concerned with high precision to ensure all visited sites are safe at the expensive of blocking some sites that were falsely identified as unsafe. Often times, the end application of the machine learning model may not only influence the performance metrics used to evaluate its generalizability, but also can inform upstream training choices. For example, to counteract class imbalance and improve the

resulting model, oversampling the minority class, under sampling the majority class, and applying class weights may be applied during training.

Overall, APA may play an important role in complex traits. Here we sought to improve methods for identifying and quantitating APA transcripts in the expressed transcriptome to enable research into their contribution to the genetics of alcohol related phenotypes.

## Statement of purpose

The major aims of this dissertation are the following:

1) **Develop a systems genetics pipeline for identifying networks of genes associated with a complex trait (Chapter II).**

   Hypothesis: Applying our method to alcohol metabolism phenotypes and liver expression data will identify candidate networks of genes that include the alcohol metabolizing genes. By taking a network approach, additional insight can be ascertained in regard to how these genes operate in the absence of alcohol and, additionally, how ingestion of alcohol may perturb these biological processes.

   Summary of approach: The systems genetics approach used three types of data: 1) genotype, 2) liver expression data (in the form of RNA-Seq), and 3) alcohol clearance and acetate area under the curve (acetate AUC) data. The RNA-Seq data and weighted gene coexpresssion network analysis (WGCNA) were used to build coexpressed gene networks. Several statistical criteria were employed to reduce the number of false positives when identifying candidate networks connecting genotype to phenotype.

2) **Develop a machine learning algorithm for identifying polyA sites in the expressed transcriptome that utilizes DNA sequence and short read bulk high-throughput RNA-Seq (Chapter III).**

   Hypothesis: An algorithm that utilizes machine learning and combines DNA sequence and RNA-Seq will outperform either single-omics method for identifying expressed polyA sites.

   Summary of approach: Short read bulk RNA-Seq, DNA sequence, and experimentally derived polyA site information were gathered from matching samples. The RNA-Seq data and DNA sequence data were used to engineer features for machine learning, and the polyA site information served as labels for supervised machine learning. A biLSTM was trained to predict polyA sites and distributed as a package written in Python.

3) **Characterize the alternative splicing and alternative polyadenylation transcriptional landscape in brain of the HXB/BXH recombinant inbred rat panel and assess its impact on predisposition to voluntary alcohol consumption (Chapter IV).**

   Hypothesis: Since at least some of the genetic component of complex traits such as voluntary alcohol consumption are likely influenced by gene regulation phenomena such as alternative splicing and alternative polyadenylation, including these transcripts in systems genetics studies on this trait will provide additional information.

   Summary of approach: A brain specific transcriptome in the HXB/BXH RI panel was generated that considered alternative splicing (via StringTie) and alternative polyadenylation (via aptardi - the algorithm developed in Aim 2). A filtering pipeline was established to identify high quality transcripts in the transcriptome. The systems genetics approach delineated in Aim 1 was applied to identify candidate transcript networks, as well as a statistical pipeline for identifying candidate individual transcripts.

# CHAPTER II

# UNSUPERVISED, STATISTICALLY-BASED SYSTEMS BIOLOGY APPROACH FOR UNRAVELING THE GENETICS OF COMPLEX TRAITS: A DEMONSTRATION WITH ETHANOL METABOLISM[1]

[1]This chapter was previously published in: Lusk R., et al., Unsupervised, statistically-based systems biology approach for unraveling the genetics of complex traits: A demonstration with ethanol metabolism. *Alcohol Clin Exp Res.* 2018;42(7):1177-1191.

## Introduction

GWAS were originally designed to leverage the principle of linkage disequilibrium at the population level by scanning millions of variants in the genome across unrelated individuals to identify loci associated with complex traits [222]. Since its first applications [223], hundreds of GWAS have been implemented and a dedicated catalog of the published studies has been developed [224, 225]. Often the QTL identified in GWAS do not fully explain the heritability of the complex trait anticipated from epidemiologic studies (e.g. alcohol dependence) [226], and the relationship between the identified loci and the biology underlying complex diseases may not be easily deciphered [227].

The advent of next-generation RNA-Seq technologies has provided researchers with new tools for gaining insight into the genetic basis of health and disease. Namely, researchers can now incorporate RNA expression levels in a "use all data" [228] systems biology approach to extract meaningful genetic information about complex traits. Integrating transcriptome expression data with genotype information, i.e. genetical genomics, can provide insight into the mechanisms underlying disease phenotypes [229]. A now standard approach for integrating information on RNA expression with genotypic information, to elucidate mechanisms by which DNA polymorphisms contribute to complex traits, is to identify the areas of the genome that are

associated with a complex trait (QTL) and that contribute to determining the levels of gene expression (expression quantitative trait loci; eQTL). Moreover, the use of methods to generate information on networks arising from analysis of gene coexpression and the genetic loci driving such coexpression (module eigengene quantitative trait loci; meQTL), can contribute additional knowledge to the underlying biology [230, 231]. For example, co-expressed genes may not only be controlled by the same transcriptional regulatory program [231], but also may be functionally linked [232]. The co-expressed gene products may be members of the same metabolic pathway or protein complex [233]. Additionally, coexpression modules can be used to functionally annotate ("guilt by association") novel or under annotated genes [234], including non-coding elements. Zhang and Horvath [235] developed a statistical technique for quantifying gene coexpression networks and identifying coexpression modules from RNA expression data. This methodology, termed weighted gene coexpression analysis (WGCNA), has been employed in numerous studies [157, 236-239] to statistically describe the relationship amongst gene products.

Our previous work has integrated expression quantitative trait loci (QTL) information and WGCNA with phenotypic QTL analysis in a hypothesis generating approach, to identify candidate modules for complex traits [132, 137, 138, 240]. In our current study, we investigated whether this approach can be valuable in a "hypothesis testing" mode. We sought to provide a proof-of-concept that an unsupervised, statistically-based systems biology approach can identify coexpression module(s) that contain well known predisposing components influencing alcohol (i.e., ethanol) metabolism. It is well-documented that the majority of alcohol (~ 95%) is eliminated via metabolism in the liver [241-243]. Hepatic oxidation of ethanol, in which alcohol dehydrogenase(s) (ADH) convert ethanol to acetaldehyde, and acetaldehyde is converted to acetate by aldehyde dehydrogenase (ALDH), is the major metabolic pathway for elimination of

ingested ethanol [243-246]. Prior studies seeking genetic explanations for differences in ethanol metabolism using QTL analysis in mice have produced results which were difficult to interpret, and did not relate to prior literature [247]. Therefore, we were attempting to validate our approach by verifying whether coexpression module(s) which were associated with alcohol clearance, derived through our analysis, reflected the extensive information in the literature on known pathways affecting alcohol clearance in previously unexposed specimens. Assuming that our analytical framework produced credible results, we further postulated that the module information that we generated would provide insight into the normal physiologic network that could be perturbed by ingestion of ethanol.

**Materials and Methods**

Unless otherwise noted, all analyses were performed using R (v. 3.3.2).

**Animals**

The HXB/BXH RI rat panel used in this study was derived from the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub) and the spontaneous hypertensive rat strain (SHR/OlaIpcv) using gender reciprocal crossing and more than 80 generations of brother/sister mating after the F2 generation [248]. While this panel was originally constructed to examine genetic control of cardiovascular phenotypes, many other complex traits have been found to vary across this panel and are thus amenable to genetic studies [129, 132, 138-142].

Male rats at the age of 90 days were used for our studies. These animals were bred and maintained at the Institute of Physiology of the Czech Academy of Sciences, Prague, Czech Republic. All experiments involving the administration of ethanol and blood sampling, as well as liver harvesting, were performed in accordance with the Animal Protection Law of the Czech

Republic and were approved by the Ethics Committee of the Institute of Physiology, Czech Academy of Sciences, Prague.

**Alcohol Clearance and Blood Acetate Level Measurements in the HXB/BXH Recombinant Inbred Rat Panel**

The alcohol experiments in the RI rat panel outlined in this section were performed specifically for this study. Namely, three male rats per strain across 30 strains of the HXB/BXH RI panel (90 rats total) were intraperitoneally injected with a 2 g/kg dose of ethanol (15% w/v). (See Supplementary Methods in Appendix S1 for a detailed description of ethanol dose choice rationale.) Blood draws from the tail vein were collected for quantifying alcohol and acetate concentration at the following time points post-alcohol administration: 20, 40, 60, 90, 120, 180, 240, 300, and 400 minutes. In addition, a 0 time point sample was gathered immediately prior to alcohol administration. For each sampling, approximately 100 µL of blood was collected. Following collection, two volumes of ice cold 0.6 N perchloric acid were added to each sample, and the resulting supernatant, after centrifugation at 13000 g and 4 °C for 10 minutes, was kept for analysis. Samples were stored at -80 °C and shipped in dry ice to the University of Colorado Anschutz Medical Campus for analysis.

**Alcohol Clearance Quantitation in the HXB/BXH Recombinant Inbred Rat Panel**

Blood alcohol levels were determined using a Varian 3800 gas chromatograph (Varian, Palo Alto, CA, USA) equipped with an Agilent Technologies DB-ALC1 column (Agilent Technologies, Santa Clara, CA, USA; part number 123-9134) and a Varian 8200 AutoSampler (Varian, Palo Alto, CA, USA). Prior to gas chromatographic analysis, the thawed blood samples were centrifuged at 13000 g and 4 °C for three minutes, and 10 µL of 100 mM 2-propanol internal standard was added to 80 µL aliquots of the supernatant. With each batch of samples

assayed, standard curves were generated using 0, 5, 10, 20, 40, 60, 80, and 100 mM alcohol standards that were prepared in blood taken from control rats in a manner identical to the experimental samples. The alcohol concentration/time curves for each rat were fit to a one-compartment pharmacokinetic (PK) model with first-order absorption and first-order elimination by employing the nonlinear Levenberg-Marquardt fitting algorithm using the minpack.lm package (v. 1.2-1) [249] in R. The first-order clearance values represent the alcohol clearance phenotype (see Supplementary Methods in Appendix S1 for a detailed description of alcohol quantitation). We used intraperitoneal injection of ethanol to avoid the confounding of metabolism of ethanol by the high $K_M$ stomach ADH system [250]. (See Supplementary Methods in Appendix S1 for a detailed description of ethanol delivery method rationale.) The pharmacokinetics of some strains of the HXB/BXH RI panel resembled pseudo zero-order alcohol elimination kinetics (e.g., BXH10 in Figure 2.1). We, however, used a uniform phenotype derived using first-order kinetic parameters to enable comparison of alcohol clearance across strains. Both statistical (Table S1) and visual comparison of zero-order (straight line) and first-order (exponential decay) kinetic models indicated little difference in fits between the two models for each animal.

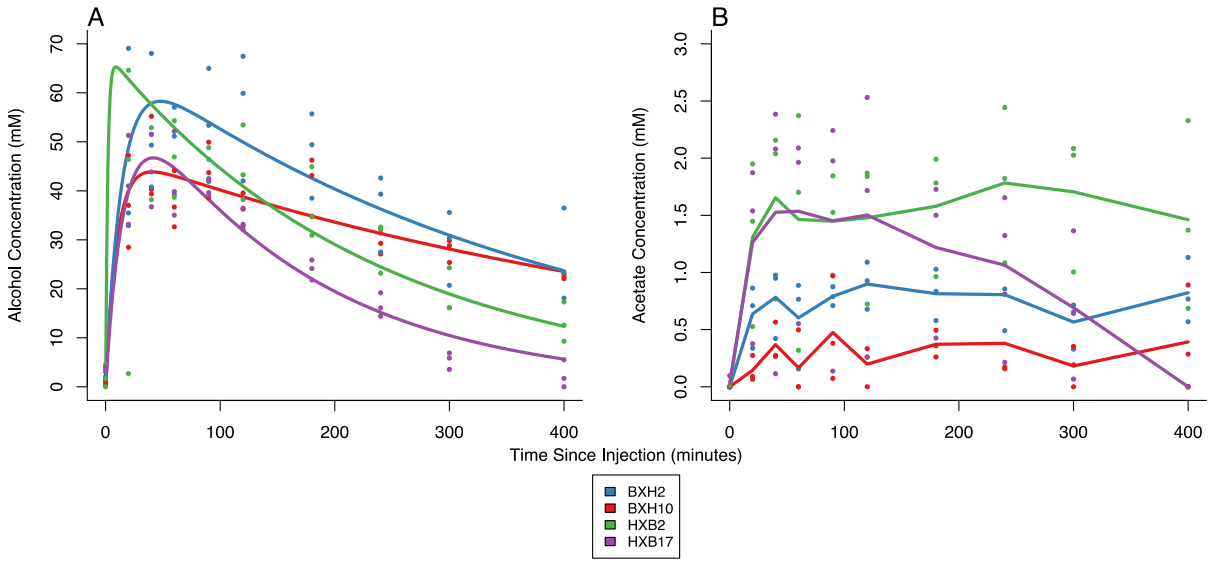



**Figure 2.1. Representative alcohol and acetate profiles in blood after 2 g/kg alcohol administration.** Concentrations in millimolar for individual animals are represented by circles at each time point for **(A) blood alcohol concentrations** and **(B) blood acetate concentrations.** The lines represent strain-specific one-compartment pharmacokinetic models with first-order absorption and elimination (a) generated from the mean of the parameter estimates from the individual rats and the lines connecting the strain mean concentrations of acetate at each time point (b).

**Acetate Area under the Curve (AUC) Quantitation in the HXB/BXH Recombinant Inbred Rat Panel**

Blood acetate levels were determined using the Sigma-Aldrich acetate colorimetric assay kit (Sigma-Aldrich, St. Louis, MO, USA; catalog number MAK086) using the manufacturer's recommended protocol. To prepare the blood samples for analysis, 48 µL of 0.5 M potassium hydroxide was added to 60 µL aliquots of the blood samples. Samples were centrifuged at 13000 g and 4 °C for one minute, and 35.7 µL of the supernatant was combined with 14.3 µL of assay buffer. For the standard curves, blood from control rats was processed in an identical manner to the experimental samples, and standard curves for concentrations of 0.00, 0.25, 0.50, 1.00, 1.50, and 2.00 mM were constructed. Absorbance values were measured using the BioTek Synergy HT plate reader (BioTek, Winooski, VT, USA). Acetate AUC from 0 to 400 minutes was

calculated from the acetate concentration-time curves for each individual rat by employing the linear trapezoidal rule using the PK package (v. 1.3-3) [251] and these values were used as the acetate AUC phenotype (see Supplementary Methods in Appendix S1 for a detailed description of acetate quantitation).

**Whole Liver RNA Sequencing for the HXB/BXH Recombinant Inbred Rat Panel**

Sequencing of livers in the RI rat panel were done for the purposes of this study. Liver tissue was stored in liquid nitrogen and shipped to the University of Colorado Anschutz Medical Campus for RNA extraction and cDNA library preparation. Total RNA extracted from livers obtained from 49 male rats (90 days old) was sequenced. Of the 49 liver samples, 44 were from the HXB/BXH RI panel (1-2 livers/strain) and five samples were from the progenitor strains (BN-Lx/Cub and SHR/OlaIpcv). The rats from these RI strains are genetically identical to the rats used for phenotyping but the rats used for RNA-Seq analysis were not exposed to alcohol. These animals were housed in identical environments as the rats that received ethanol.

Livers were processed in three batches and included seven technical replicates (56 libraries). Total RNA (>200 bases) was extracted and cleaned using the RNeasy Plus Universal Midi Kit and Rneasy Mini Kit, respectively (Qiagen, Valencia, CA, USA). Four µL of a 1:100 dilution of either ERCC Spike-In Mix 1 or Mix 2 (ThermoFisher Scientific, Wilmington, DE, USA) were added to each extracted RNA sample. Construction of sequencing libraries was done using the Illumina TruSeq Stranded RNA Sample Preparation kit (Illumina, San Diego, CA, USA) in accordance with the manufacturer's protocol. Part of this process included ribosomal RNA depletion via the Ribo-Zero rRNA reduction chemistry. An Agilent Technologies Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA) was utilized to assess sequencing library quality, and samples were sequenced (2x100 paired-end (PE) reads, three to

four samples multiplexed per lane) on an Illumina HiSeq2500 (Illumina, San Diego, CA, USA) in High Output mode.

**Quantitation of Whole Liver RNA for the HXB/BXH Recombinant Inbred Rat Panel**

Raw reads were trimmed to remove adapter sequences as well as low quality bases using Trim Galore! (v. 0.4.0). Low quality bases were determined using the default parameters. The trimmed reads were initially aligned to ribosomal RNA (rRNA) from the RepeatMasker database [252] accessed through the UCSC Genome Browser [253, 254] using TopHat (v. 2.0.14) [255]. PE reads which did not map to rRNA were quantified into Ensembl gene-level abundance estimates (Ensembl Release 81) [256] using the RSEM (RNA-Seq by Expectation Maximization package) (v. 1.2.21) [257] and strain-specific Ensembl transcriptomes generated in our laboratories. (See Supplementary Methods in Appendix S1 for a detailed description of Trim Galore!, TopHat, and RSEM settings). Initially, strain-specific genomes for the RI strains were constructed from the Rat Genome Sequencing Consortium (RGSC) Rnor_6.0 version of the rat genome [258] by imputing single nucleotide polymorphism (SNP) information for each strain based on their STAR Consortium genotypes [259] and DNA sequencing (DNA-Seq) data from male rats of the progenitor strains [86]. These data are publicly available on the PhenoGen website [132, 260]. Strain-specific transcriptomes were generated from these imputed genomes and the Ensembl database (Ensembl Release 81) [256].

To prepare the expression estimates for analysis, genes with an average RSEM-estimated read count of less than one across the 56 samples were considered undetectable above background and not used in the analysis. Quantitated samples were initially examined for quality using hierarchical clustering. The RUV (Removal of Unwanted Variance) algorithm [261] based on empirically-derived control genes was used to eliminate batch effects and other technical

factors contributing to variance. Empirically-derived control genes were identified as the 5,000 least significant genes in a negative binomial generalized linear model [262] with RI strain as the covariate using the edgeR package (v. 3.14.0) [263] in R. The RUVg function from the RUVSeq package (v. 1.6.2) [261] was used to derive three normalization factors. The number of normalization factors used was determined by the clustering of technical replicates. Normalized counts were used in subsequent analyses.

After normalization, samples were reduced to only include RI animals (not the progenitors) and only one technical replicate per biological sample, i.e. multiple measurements from the same animal performed for analysis of unwanted variance were reduced to a single measurement (the replicate with the highest normalized read count was retained). Because of the reduction in samples and the normalization of read counts, genes were again filtered based on the criteria of an average normalized read count greater than one. The normalized expression data were transformed into regularized log (rlog) values using the DESeq2 package (v. 1.12.4) [264]. This function 1) transformed the data to a $\log_2$ scale and 2) stabilized the within gene variance to avoid the dependence of the variance on the mean.

**Weighted Gene Co-Expression Network Analysis for the HXB/BXH Recombinant Inbred Rat Panel**

The WGCNA package (v. 1.51) [265] was used to build coexpression modules from the rlog-transformed RNA expression estimates for the HXB/BXH RI panel collected for this study. For HXB/BXH RI strains with multiple samples, i.e. measurements on independent animals but of the same strain, mean strain values were used to calculate connectivity. Two settings of the network-building function for WGCNA were changed from their default settings: the minimum module size parameter (set to five instead of 30) and the deepSplit parameter (set to four instead

of two). Both of these alterations promote the identification of smaller modules, which was desirable because it facilitates subsequent independent expert analysis of how genes contained within any identified candidate modules are interrelated for ascertaining biological insights. The potential trade-off between neglecting additional genes that otherwise would have been included in a given module for manual interpretation is mitigated by the fact that coexpression between modules can be elucidated. The absolute value of the Pearson correlation coefficient was used to determine the adjacency matrix, i.e. an unsigned network. Furthermore, the soft-thresholding index, β, was set to six to approximate a scale-free topology. The value for the soft-thresholding index was determined using the methods and critical values proposed in Zhang and Horvath [235] (Figure S1). A module eigengene (first principal component) was used to summarize the gene expression profiles within a module across strains for subsequent analyses [265].

**Quantitative Trait Loci Analysis (QTL)**

The molecular marker set used for QTL analyses was derived from our existing publicly available strain-specific genomes created using SNPs genotyped by the STAR Consortium [259]. Probes from the arrays used to generate this marker set were aligned to the RGSC Rnor_6.0 rat genome assembly [258, 266] using the UCSC command line BLAST-like alignment tool (BLAT) [267], and both the genome and the alignment tool were downloaded from the UCSC Genome Browser [253, 254]. We retained markers for QTL analysis in this study if the following criteria were met: 1) their probe sequence aligned perfectly and uniquely to the genome, 2) their genotypes differed between progenitor strains, 3) neither progenitor strain was heterozygous for the SNP, and 4) less than 5% of the HXB/BXH RI strains were missing or heterozygous for the SNP. In addition, markers with large estimated genetic distance compared to physical distance from adjacent markers (improbable recombination events, flanked by more than 10 cM on each

side) and double recombinant markers were removed. Genetic distances were estimated using the R/qtl package (v. 1.40-8) [268]. Prior to QTL analyses, the marker set was reduced to unique strain distribution patterns, i.e., multiple adjacent markers with the same genotype pattern across strains were represented by a single marker, in order to reduce the computational burden.

Marker regression was used to calculate module eigengene QTL using strains of the HXB/BXH RI panel that had both genotype information previously acquired by our group and RNA expression/eigengene estimates gathered for this study. The pQTL for alcohol clearance and acetate AUC in the HXB/BXH RI panel were also determined using marker regression with strain means. Empirical genome-wide p-values were calculated using 1000 permutations [269]. For the two phenotypes, both significant ($p < 0.05$) and suggestive ($p < 0.63$) pQTL were considered. The definitions of significant and suggestive p-values were taken from Lander and Kruglyak [270] and have been adopted by others [271]. The 95% Bayesian credible interval of each meQTL and pQTL was calculated using the methods detailed in Sen and Churchill [272]. All QTL analyses and graphics were generated using the R/qtl package (v. 1.40-8) [268].

**Identification of Candidate Modules for Alcohol Clearance and Acetate Area under the Curve in the HXB/BXH Recombinant Inbred Rat Panel**

The first step for identifying candidate coexpression modules for alcohol clearance and acetate AUC was to evaluate their association with each phenotype. Strain mean values of alcohol clearance and acetate AUC procured for this study were used for correlation analysis. A Pearson correlation coefficient between the module eigengene and the phenotype across the strains of the HXB/BXH RI panel was estimated for each module and phenotype. Only modules significantly associated (nominal p-value < 0.01) with at least one of the two phenotypes were considered in subsequent steps.

For modules correlated with alcohol clearance and/or acetate AUC, additional criteria were imposed using our previously obtained QTL data in order to be considered a candidate module for either phenotype. Candidate modules were required to have a genome-wide significant (p-value < 0.01) meQTL, and the module meQTL must fall within the 95% Bayesian credible interval of a significant or suggestive pQTL for the given phenotype.

## Results

### Alcohol Clearance and Acetate Area Under the Curve in the HXB/BXH Recombinant Inbred Rat Panel

Representative blood alcohol and acetate profiles (Figure 2.1) demonstrate the diversity of the blood alcohol and acetate profiles across the HXB/BXH RI panel. Overall, 82 rats and 691 measurements were used for alcohol clearance calculations after quality control. (See Supplementary Results in Appendix S1 for detailed results from quality control imposed on alcohol and acetate measures). Average alcohol clearance varied approximately 10-fold among strains in the RI rat panel (0.8 to 7.5 mL/min/kg; Figure 2.2A). Furthermore, the panel exhibited high broad-sense heritability (81%) for this phenotype, estimated as the coefficient of determination from a one-way ANOVA. After quality control, 89 rats and 888 measurements were used for acetate AUC calculations. Peak circulating blood acetate levels varied from 0.20 to 2.74 mM (interquartile range: 0.84 to 1.79 mM) among strains, and acetate (AUC) varied from 82 to 617 mM*min and displayed a high broad-sense heritability (66%; Figure 2.2B). Using strain means, alcohol clearance and acetate AUC were positively correlated (Pearson's correlation coefficient = 0.43, 95% CI = 0.09 to 0.69, p-value = 0.016; Figure 2.2C).

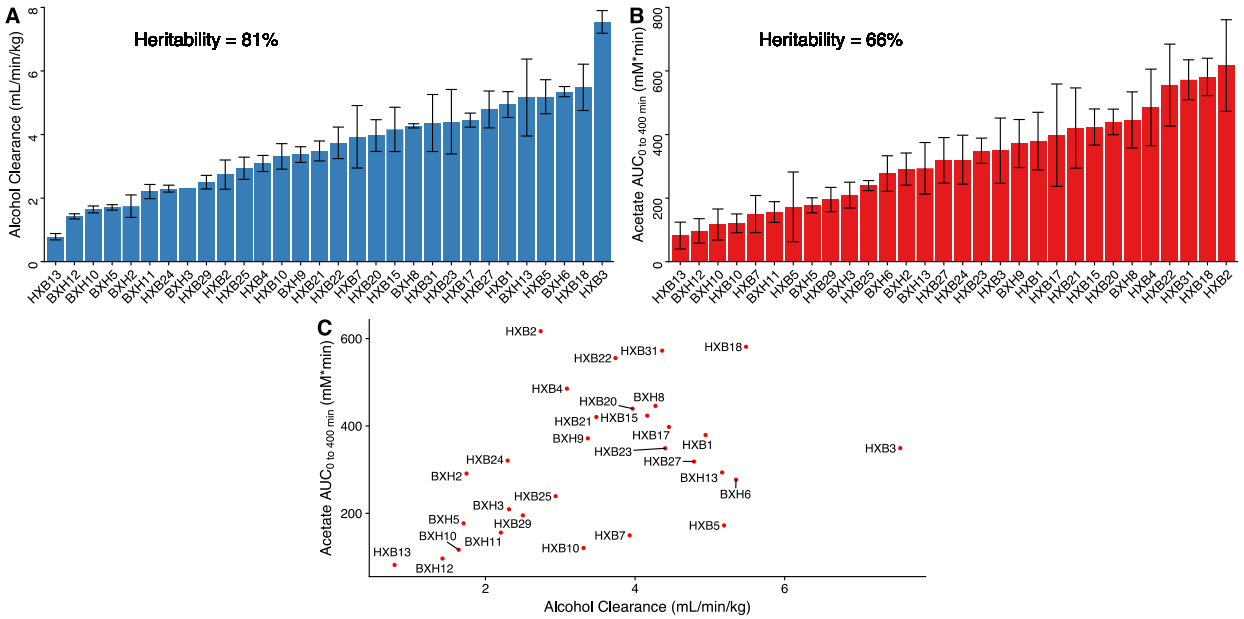



**Figure 2.2. Distribution of alcohol clearance and acetate AUC across the HXB/BXH recombinant inbred rat panel.** The bars represent mean values of the biological replicates within the strain denoted on the x-axis for **(A) first-order alcohol clearance** (blue) and **(B) acetate AUC** (red). The error bars represent plus/minus standard error of the mean. If error bars are missing, only one biological replicate was available for the given strain. Alcohol clearance estimate and acetate AUCs were determined in each rat separately. The broad sense heritability of each phenotype was estimated as the R-squared value from a one-way ANOVA using strain as the predictor. Mean values for the two phenotypes were plotted against each other by strain to examine the **(C) association between alcohol clearance and acetate AUC**. Each point is labeled by its respective strain.

## Whole Liver RNA Sequencing for the HXB/BXH Recombinant Inbred Rat Panel

RNA-Seq was performed on RNA extracted from livers of naïve (non-alcohol exposed) rats in three batches (56 libraries including technical replicates). Over three billion total paired-end (PE) reads were generated from these samples. This amounts to approximately 60 million PE reads per sample. After trimming and removal of reads that aligned to ribosomal RNA (rRNA), the average number of PE reads per sample was 59.5 million and 58.4 million, respectively. A detailed summary of the RNA-Seq results by sample is in Table S2.

**Quantitation of Whole Liver RNA for the HXB/BXH Recombinant Inbred Rat Panel**

We eliminated Ensembl genes with an average estimated RSEM count of less than one across the 56 rat liver RNA-Seq libraries. This resulted in a reduction from 32285 to 16093 Ensembl genes. Based on visual inspection of the Pearson correlation between samples using $\log_2$(RSEM counts + 1) transformed data, four samples were identified as outliers and removed (Figure S2). The dendrograms of the samples (including information on technical replicates and batches) before and after implementation of the RUV algorithm provided evidence that unwanted variance, such as that introduced by batch effects, was markedly reduced. Removal of data from progenitor strains and technical replicates (used for normalization) left data from 41 HXB/BXH RI samples (1-2 rats/strain; 29 strains), and the further removal of genes with normalized counts less than this final number of samples left 15984 Ensembl gene identified in liver. These data were utilized in WGCNA.

**Weighted Gene Co-Expression Network Analysis for the HXB/BXH Recombinant Inbred Rat Panel**

For strains in which RNA-Seq data were obtained (29 HXB/BXH RI strains), the expression estimates were subjected to WGCNA to identify coexpression modules. A total of 658 modules were identified (median module size = 8 genes; Figure S3) along with 205 genes that could not be assigned to a module. The module eigengenes captured much of the within-module variability in expression across strains (interquartile range: 59% to 67%).

**Quantitative Trait Loci Analyses**

A total of 20283 SNPs were originally contained in the STAR dataset [259]. After processing (see Supplementary Results in Appendix S1 for detailed results from processing), we identified 1529 unique strain distribution patterns, i.e., haplotype blocks, for the 32 HXB/BXH

RI strains genotyped by the STAR Consortium. Of 32 HXB/BXH RI strains with genotype information, 29 strains had expression/eigengene estimates and were used to calculate meQTL, and 30 strains had alcohol clearance and acetate AUC data and were used for pQTL analysis. The HXB21 RI strain had alcohol clearance and acetate AUC data and was therefore used in pQTL analysis, but the RNA-Seq data were removed as outliers and therefore were not used in meQTL analysis. The pQTL analysis using 1,000 permutations identified one significant (genome-wide p-value < 0.05) and two suggestive (genome-wide p-value < 0.63) pQTL for alcohol clearance (Figure 2.3A) and four suggestive pQTL for acetate AUC (Figure 2.3B).

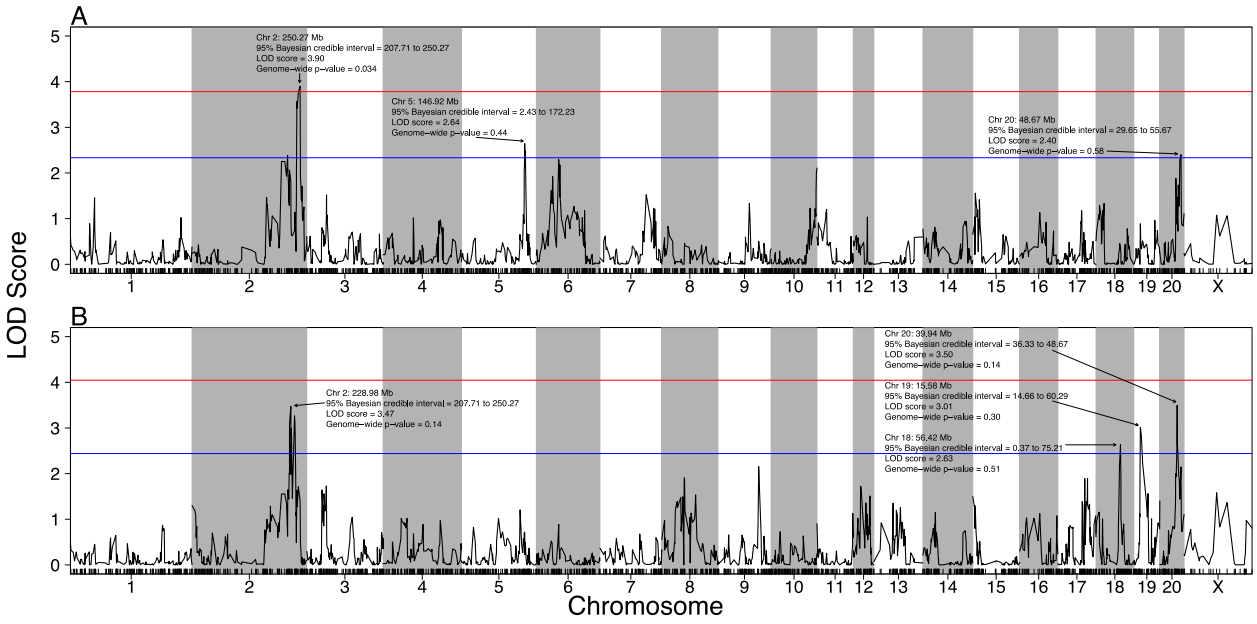



**Figure 2.3. Quantitative trait loci for alcohol clearance and acetate AUC in the HXB/BXH recombinant inbred panel.** Strain means were used in a marker regression to determine phenotypic QTL for **(A) alcohol clearance** and **(B) acetate AUC.** The red lines represent the logarithm of odds (LOD) score threshold for a significant QTL (genome-wide p-value = 0.05), and the blue lines represent the LOD threshold for a suggestive QTL (genome-wide p-value = 0.63). Significant and suggestive QTL are labeled with their location, 95% Bayesian credible interval, LOD score, and genome-wide p-value.

**Identification of Candidate Modules for Alcohol Clearance and Acetate Area under the Curve in the HXB/BXH Recombinant Inbred Rat Panel**

RNA expression data from alcohol-naïve rats were used to identify the transcriptional predisposing factors for alcohol clearance and circulating acetate levels after administration of ethanol. Module eigengenes were correlated with strain mean values of alcohol clearance and acetate AUC separately. Ten modules were significantly (nominal p-value < 0.01) correlated with alcohol clearance, and ten modules were significantly associated with acetate AUC; moreover, three modules were correlated with both phenotypes (Table S3). The same marker set used for pQTL analyses was used to identify the meQTL with the greatest logarithm of odds (LOD) score for each module that had a significant correlation with either alcohol clearance or acetate AUC. Of these modules, one module associated with alcohol clearance had a significant (genome-wide p-value < 0.01) module eigengene QTL. The examination of overlap between the 95% Bayesian credible intervals of the alcohol tolerance pQTL and the location of the peak LOD score for the module eigengene QTL demonstrated that only this candidate module (orange3) met all the criteria to be identified as a candidate module influencing alcohol clearance. This module was also identified as a candidate module for acetate AUC. The genes comprising the orange3 module are listed in Table 1.1, and the connectivity between genes of the candidate module is visualized in Figure 2.4.

**Table 2.1. Genes comprising the orange3 candidate module for both alcohol clearance and acetate AUC.**

| Associated Gene Name | Chromosome | Gene Starting Position (Mb) | Type | Description | Intra-modular Connectivity |
|---|---|---|---|---|---|
| *Adh4* | 2 | 243.70 | Protein coding | Alcohol dehydrogenase 4 (class II) | 1.18 |
| *Adh1* | 2 | 243.55 | Protein coding | Alcohol dehydrogenase 1 (class I) | 1.17 |
| *Hs2st1* | 2 | 250.47 | Protein coding | Heparan sulfate 2-O-sulfotransferase 1 | 1.14 |
| *Arl16* | 10 | 109.64 | Protein coding | ADP-ribosylation factor-like 16 | 1.13 |
| *Gbp5* | 2 | 248.18 | Protein coding | Guanylate binding protein 5 | 1.11 |
| *Camk2n1* | 5 | 156.88 | Protein coding | Calcium/calmodulin-dependent protein kinase II inhibitor 1 | 1.11 |
| *LOC685067* | 2 | 248.22 | Protein coding | Similar to guanylate binding protein family, member 6 | 1.10 |
| *Tmem79* | 2 | 187.70 | Protein coding | Transmembrane protein 79 | 1.08 |
| *Zfp143* | 1 | 174.70 | Protein coding | Zinc finger protein 143 | 1.07 |
| *Piwil2* | 15 | 52.04 | Protein coding | Piwi-like RNA-mediated gene silencing 2 | 1.05 |

The candidate module was identified based on eigengene association (nominal p-value < 0.01) with the phenotypes across the HXB/BXH panel and a maximum module eigengene quantitative trait locus that was both significant (genome-wide p-value < 0.01) and overlapped significant (genome-wide p-value < 0.05) or suggestive (genome-wide p-value < 0.63) phenotypic quantitative trait loci for the phenotypes in the HXB/BXH RI rat panel after two g/kg alcohol administration. Genes are ordered by intra-modular connectivity.

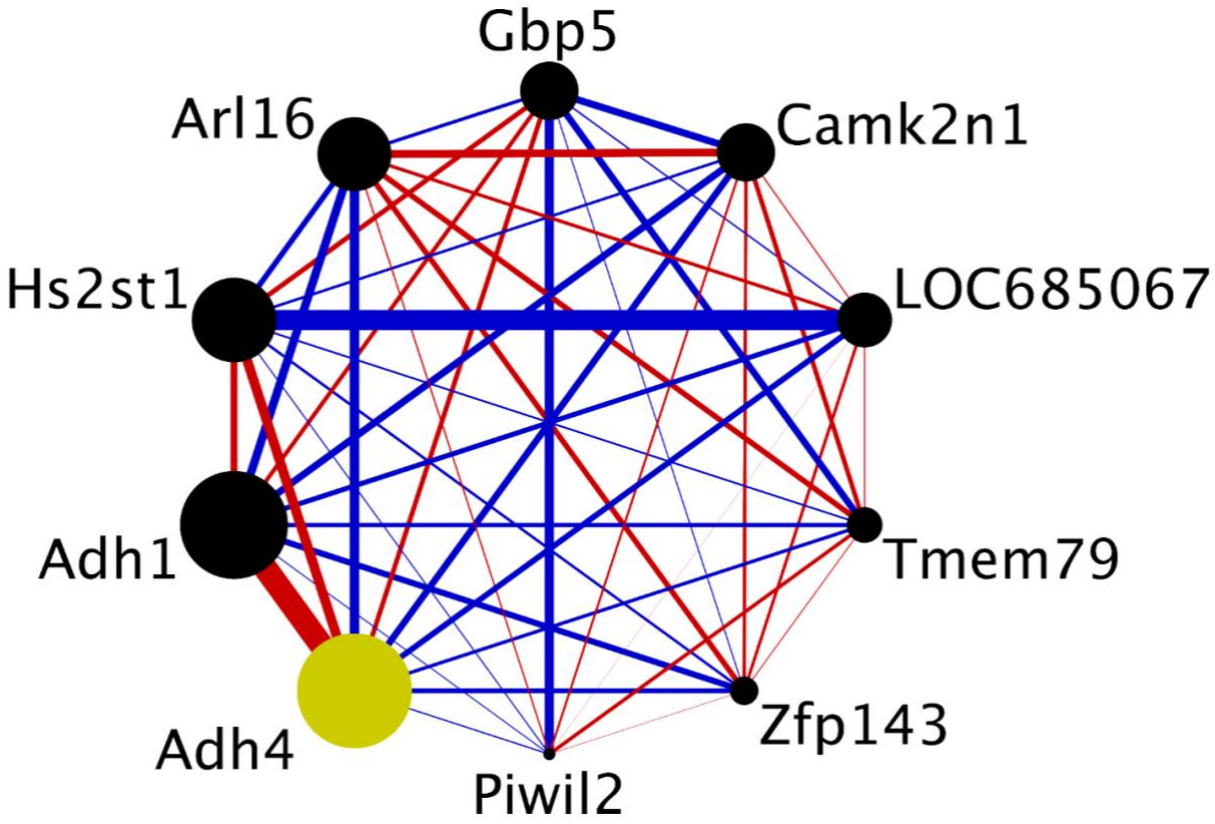



**Figure 2.4. Connectivity within the candidate coexpression module for both alcohol clearance and acetate area under the curve (AUC) after two g/kg alcohol administration.** The module eigengene is significantly ($p < 0.01$) correlated with alcohol clearance and acetate AUC. Furthermore, its module eigengene QTL is significant ($p < 0.05$) and overlaps a phenotypic QTL for alcohol clearance and acetate AUC, respectively. Each node represents a gene from the coexpression module. The size of each node is weighted based on its intra-modular connectivity, and the thickness of each edge is weighted based on the magnitude of the connectivity between the two genes. The edge colors indicate the direction of the connectivity (red = positive, blue = negative). The hub gene, defined here as the single gene with the largest intra-modular connectivity, is colored in yellow (Adh4; alcohol dehydrogenase 4), and its expression is positively associated with both alcohol clearance and acetate AUC. The figure was generated using Cytoscape (v. 3.4.0) [273].

The module eigengene for orange3 explained 56% of the genetic variance in alcohol clearance and 32% of the genetic variance in acetate AUC across the HXB/BXH RI rat panel, estimated as the coefficient of determination from a linear model using the module eigengene as a predictor of each phenotype. The hub gene, i.e. the gene with the highest intra-modular connectivity within the orange3 candidate module, was alcohol dehydrogenase 4 (*Adh4*).

**Characterization of Alcohol Dehydrogenase Genes in the orange3 Candidate Module**

Two alcohol dehydrogenase Ensembl genes, alcohol dehydrogenase 6 (*Adh6*) and *Adh4*, were initially identified in the orange3 candidate module. We further examined these genes at the transcript-level to verify their identities. The *Adh6* gene expression estimate was found to be derived from pooled estimates of three Ensembl transcripts. Using the University of California – Santa Cruz (UCSC) Genome Browser [253, 254], we found that the RefSeq database annotated two of these transcripts as separate genes: namely *Adh6* and the class I alcohol dehydrogenase 1 (*Adh1*). The remaining transcript was unannotated in RefSeq, and closer inspection in Ensembl (Ensembl Release 88) [274] revealed that this transcript represented a fusion of *Adh1* and *Adh6*. To disentangle the two *Adh* genes, we first examined the pile-up of the total RNA-Seq reads from the livers of the two progenitor strains. The vast majority of reads aligned to the RefSeq *Adh1* gene, with very few reads aligning to *Adh6,* and there was no evidence for expression of the fusion gene (Figure S4A). To verify that the variation in expression levels of the original gene-level estimate mimic variation in *Adh1*, pairwise Pearson's correlation analysis was performed between the Ensembl gene-level expression estimate, Ensembl transcript-level expression estimates of the three transcripts, and the phenotypes, using strain mean values. Expression estimates of the transcripts were calculated in an identical manner as the gene-level estimate, i.e., rlog-transformed batch-corrected RNA expression values. The expression of the Ensembl transcript corresponding to the RefSeq *Adh1* gene was most strongly correlated with the overall Ensembl *Adh6* gene expression (Pearson's r = 0.95) and closely matched its correlation with the phenotypes (Table S4). Moreover, the Ensembl database used for our initial annotation (Ensembl Release 81) [256] did not annotate any transcripts as *Adh1*, in spite of the fact that RNA from *Adh1* is known to be present in rat liver [275]. Therefore, we concluded that the gene-

level expression estimates originally annotated as *Adh6* in fact most likely represented *Adh1* and changed the annotation throughout this manuscript accordingly.

Likewise, *Adh4* included pooled estimates from three Ensembl transcripts. Again, two of the Ensembl transcripts were annotated by RefSeq as separate genes, *Adh4* and class III alcohol dehydrogenase 5 (*Adh5*), and the third represented a fusion of these genes that was unannotated in other databases. In this case, the expression of the Ensembl transcript corresponding to the RefSeq *Adh4* gene was most closely correlated with the overall Ensembl *Adh4* gene-level expression (Pearson's r = 0.97) and most closely resembled its correlation with the phenotypes (Table S4). While the liver total RNA-Seq reads from the progenitor strains mapped to both the *Adh4* and *Adh5* RefSeq genes, indicating that both were expressed in the liver (Figure S4B), the Ensembl transcript/RefSeq *Adh4* gene demonstrated greater variation in expression across strains. Taken together, we surmised *Adh4* was indeed the gene represented in the orange3 candidate module, and the original nomenclature was retained. Furthermore, BLAST analysis [276] revealed that the Ensembl transcript/RefSeq *Adh4* that we identified shared > 99% sequence similarity with the experimentally cloned and sequenced class II rat *Adh4* gene [277], thereby supporting the identity of the transcript that we sequenced as the class II alcohol dehydrogenase gene product in rat.

**Further Review of Alcohol Dehydrogenase Genes, Aldehyde Dehydrogenase Genes, and Other Genes That Can Contribute to Alcohol Metabolism in the HXB/BXH RI Panel**

All Ensembl annotated alcohol dehydrogenase genes and aldehyde dehydrogenase genes with expression estimates, as well as the genes encoding catalase (*Cat*) and Cytochrome P450 2E1 (*Cyp2e1*), were examined to ascertain their expression and determine the correlation of their expression estimates with alcohol clearance and acetate AUC in the HXB/BXH RI panel (Table

S5). Using pairwise Pearson correlation analysis on strain mean values, significant (nominal p-value < 0.05) associations were only found between the following: expression of *Adh1* and alcohol clearance (Pearson's r = 0.76, p < 0.001), expression of *Adh1* and acetate AUC (Pearson's r = 0.57, p-value = 0.0012), expression of *Adh4* and alcohol clearance (Pearson's r = 0.64, p-value = 0.0002), and expression of *Adh4* and acetate AUC (Pearson's r = 0.52, p-value = 0.0037). However, we noted that the correlation between expression of *Aldh1a1* (aldehyde dehydrogenase 1 family, member A1) and acetate AUC was marginally significant (Pearson's r = 0.36, p-value = 0.052). If we accept the premise that expression levels of *Aldh1a1* are contributing to the variation of acetate AUC in the panel, then *Aldh1a1* transcript levels explain approximately 13% of the variance in acetate AUC.

## Discussion

### Candidate Module for Alcohol Clearance – orange3

Of the 658 coexpression modules built from the liver "total" RNA-Seq data across naïve rats of the HXB/BXH RI panel, only one module satisfied all criteria as a candidate module for alcohol clearance. The orange3 candidate module contained two alcohol dehydrogenase transcripts**,** *Adh1* and *Adh4*, which produce the alcohol dehydrogenase enzymes class I Adh1 and class II Adh4**,** respectively. Adh1 is a high affinity, i.e., low $K_m$, enzyme for alcohol (~ 1.4 mM) that is mainly expressed in liver, where it accounts for the majority of alcohol elimination in rats [278, 279]. Adh4 in rat is analogous to the human variant [280] and is also expressed in the liver [281, 282]. While the $K_m$ value of rat Adh4 may be greater than that of its human counterpart [282], similar to human ADH4, the rat enzyme most likely contributes to the metabolism of alcohol in the liver at higher concentrations of alcohol [245]. Two other liver alcohol dehydrogenase genes in the rat have been reported in the literature – class III alcohol

dehydrogenase 5 (*Adh5*) and class IV alcohol dehydrogenase 7 (*Adh7*) [283]. The Rnor_6.0 version of the rat genome lacked any Ensembl annotation for *Adh5*; however, RefSeq annotation indicates that one of the Ensembl transcripts quantitated in our RNA-Seq data may actually represent *Adh5*. Indeed, research has demonstrated that *Adh5* is expressed in the rat liver [280], and the pile-up of liver total RNA-Seq reads from the progenitor strains indicates significant levels of *Adh5* expression. Nevertheless, the gene product of *Adh5* is believed to have no detectable ethanol metabolizing activity at concentrations reached in our studies [278, 280, 283]. An additional alcohol dehydrogenase gene, *Adhfe1*, existed in Ensembl annotation but is known for metabolizing 4-hydroxybutyrate in mammals rather than alcohol [284]. Finally, expression of both human and rat *ADH7*/*Adh7* has been established as exclusive to the stomach [285]. Our findings corroborated this view, as we found little *Adh7* expression in the liver (Figure S6). Overall, we were satisfied that our unsupervised statistically-based systems biology approach could clearly reproduce an accepted fact about the importance of Adh1 and Adh4 in the metabolism of alcohol when alcohol is present at levels attained in our studies.

The contradictory information regarding Ensembl and RefSeq annotations (see Results) with regard to *Adh1* and *Adh4*, however, highlights both the need to carefully examine the results obtained from high throughput RNA-Seq analyses, and the intrinsic advantages of next-generation sequencing technologies like RNA-Seq over methods such as microarrays. Namely, RNA-Seq allows one to examine *post-hoc* where reads aligned to the genome, and accordingly make annotation adjustments as necessary. Indeed, updated Ensembl versions, for example Ensembl Release 88 [274] and newer, changed the *Adh6* gene annotation in the Rnor_6.0 rat genome to *Adh1* in agreement with the annotation used throughout this manuscript.

**Characterization of Other Genes in the orange3 Module and Common Genetic Pathways**

The added benefit of a systems biology approach is that it provides biologic context for the statistically-derived relationships between transcripts contained in a module. Gene products composing the orange3 candidate module are listed in Table 1.1. While products of *Adh4* and *Adh1* are well known for their ethanol metabolizing function, alcohol dehydrogenases also metabolize a wide variety of other substances, such as longer chain aliphatic alcohols [286, 287], omega-hydroxy-fatty acids [277, 288], hydroxysteroids [277], and lipid peroxidation products [288-290]. Another substrate for ADH enzymes is all-trans-retinol which is the alcohol form of vitamin A [291]. The active metabolite of vitamin A is retinoic acid and the initial step in the conversion of the retinol to retinoic acid is catalyzed by Adh1 as well as members of the retinol dehydrogenase families [291]. The importance of Adh1 in this metabolic step is demonstrated by the knock-out of this enzyme, which results in accumulation of retinol in adult mice, and a greater retinol toxicity in the adult tissues [291]. The role of Adh1 in retinol metabolism provides one of the links to explain the association of Adh and the other gene products in the orange3 module. Retinoic acid plays a number of physiological roles through binding to cellular retinoic acid receptors (RARs) that control transcription [292]. With regard to the components of the orange3 module, the induction of the retinoic-acid inducible gene I (RIG-I) is of interest [293]. RIG-I is a helicase which functions to destroy a number of RNA viruses that may enter the cell. The RIG-I pathway is tightly regulated to maximize antiviral immunity while minimizing immune-related pathology. The product of the orange3 module member ADP-ribosylation factor-like 16 (*Arl16*) is a protein that interacts with RIG-I and inhibits its activity. Arl16 is part of the extended ADP ribosylation factor (ARF) family of GTPases, and although the ADP-ribosylation factor-like (ARL) proteins have actions beyond those exhibited by the ARF

GTPases, they also participate in the regulation of secretion, phagocytosis, endocytosis and signal transduction characteristic of the ARF GTPases [294]. RIG-I activation also leads to the production of interferon (IFN), which in turn is the major inducer of transcription of guanylate-binding proteins (GBPs). The guanylate binding protein 5 (*Gbp5*) gene is a member of the orange3 module, and synthesis of its protein (Gbp5) is responsive to IFN-γ [295]. Gbp5 is a member of the dynamin family of GTPases and recent studies have shown it to be a critical factor in the assembly of inflammasomes. Overexpression of Gbp5 enhances the expression of IFN and other pro-inflammatory factors [296], which generates a feed forward immune response and antiviral activity [297]. Since *LOC685067* is an under-annotated gene described as "similar to guanylate binding protein family, member 6", the fact that it shares membership with another guanylate binding protein in the orange3 module may add rationale to its description.

*Hs2st1* (heparin sulfate 2-O-sulfotransferase) encodes a member of the heparin sulfate biosynthesis pathway [298]. Heparin sulfate is part of a family of heparin/heparin sulfate glycosaminoglycans that organize at the cell surface to act as recognition and binding sites for chemokines [299, 300], transforming growth factors [301], and viruses [302, 303]. It should be noted that the quantity and location of sulfate groups on the heparin sulfate polysaccharide is a determining factor in the selectivity of the cell surface polyglycan for various ligands [304, 305]. Acetylation of heparin sulfate is a further modification of the glycosaminoglycans and is important in determining the recognition of various chemokines [305]. We postulate that the production of acetate from ethanol could influence this molecular modification. On the whole, the production and modification of heparin sulfate is an important component of both the capacity of pathogens, particularly viruses, to infect cells, and for the cell to mount an immune response to the pathogen. There also exists an under-investigated interaction between the

heparin/heparin-sulfate glycosaminoglycans generated by the actions of the Hs2st1 protein and another member of the orange3 module. Heparin has been shown to inhibit phosphorylation and the generation of autonomous activity of the calcium/calmodulin-dependent protein kinase II (CaMKII) [306]. The protein product of the module member calcium/calmodulin-dependent protein kinase II inhibitor (*Camk2n1*) is also an inhibitor of CaMKII. CaMKII is a multipurpose calcium/calmodulin signal transduction enzyme, best known for its role in generating cellular memory (specifically, long-term potentiation) in the hippocampus. In relation to the liver and the orange3 module, CaMKII has an important role in controlling tumor necrosis factor alpha (Tnf-α)-induced expression of CD44 [307]. CD44 is a transmembrane glycoprotein expressed in many cell types. A key event in the activation of monocytes and their transformation, cytokine release, and migration to sites of inflammation and tissue injury is the induction of CD44 expression [307]. Interestingly, ethanol is known to increase circulating levels of Tnf-α, and the *Camk2n1* gene product may be a modifier of this response.

*Zfp143* (zinc finger protein 143) encodes a zinc finger transcriptional regulator. Its human counterpart, *ZNF143* (zinc finger protein 143), exhibits the interesting property of connecting promoter regions of DNA with distant regulatory elements through looping of chromatin [308], and the protein product of *ZNF143* has been hypothesized to influence differentiation and cell identity [308]. The *Piwil2* (piwi like RNA-mediated gene silencing 2) gene product has previously been shown to be expressed in liver and is considered to be involved in regeneration of liver after damage [309]. The function of the product of the *Tmem79* (transmembrane protein 79) gene is not known with regard to liver, but its inclusion in this module may provide some insights.

The module as a whole provides the impression that in a basal state (without having substantial amount of ethanol in the milieu) the co-expressed genes are functioning as components which contribute to cell interaction with pathogens (possibly viruses), cellular response to pathogens and immune system signals, and components of liver regeneration (if it sustains damage). The alcohol dehydrogenases included in this module may be the enzymes that generate the necessary ligands (e.g., retinoic acid or acetate) critical for the function of the other module components. The association of the alcohol dehydrogenase genes with the other module components may also indicate cross-cell type communication, with retinoic acid production in hepatocytes being utilized for function of other liver cell types (e.g., Kupffer cells and/or infiltrating macrophages).

The relationship of the alcohol dehydrogenase-containing orange3 module with immune function may be through promoting what is called the autonomous immune response directed at viral infections. The chronic consumption of alcohol in excess of 50 g/day increases an individual's vulnerability to the hepatitis C virus (HCV) [310]. McCartney et al. [311] produced evidence that ethanol metabolism, rather than ethanol *per se*, promotes the replication of HCV and diminishes the antiviral action of interferon $\alpha$. Our data may contribute to the interpretation of mechanisms by which ethanol metabolism promotes the development of viral hepatitis.

The orange3 module was additionally identified as a candidate module for influencing acetate AUC. This result indicates that the rate of alcohol clearance influences acetate AUC, i.e. systemic blood acetate levels are at least partially determined by the rate at which alcohol is metabolized. Such an observation logically follows what has been reported in the literature; that is, the majority of alcohol is cleared via oxidative metabolism [242, 243] that generates acetate [245].

A somewhat surprising outcome was that, although the alcohol dehydrogenase-containing module contributed to the determining circulating levels of acetate, enzymes essential for conversion of acetaldehyde to acetate (aldehyde dehydrogenases) were not identified through the process of coexpression module analysis as being responsible for circulating levels of acetate. A cogent explanation of this fact would be, that under our experimental conditions, the aldehyde dehydrogenases that catalyze an irreversible production of acetate are not rate-limiting in the transition of ethanol to acetate. On the other hand, not all expressed genes can be assigned to coexpression modules, or even if assigned to a module, the module eigengene values for that module may not generate a statistically significant meQTL. With regard to *Aldh1a1*, we noted that it was included in an identified module but that the module's eigengene values did not generate a meQTL which overlapped the pQTL for acetate AUC. On the other hand, an eQTL for the expression levels of *Aldh1a1 per se* was associated with a SNP located at chr20: 44.71 Mb (p-value = 0.006 via permutation on PhenoGen) and overlapped the pQTL for acetate AUC (chr20: 39.94 Mb, 95% Bayesian credible interval = 36.33-48.67 Mb, LOD score = 3.50). The expression levels of *Aldh1a1* did, also, nominally correlate with the acetate AUC values although no correlation was evident with values for ethanol clearance. These results indicate to us that under some circumstances, transcript products may act outside of the context of the module to which they belong in order to carry out metabolic functions not normally part of the repertoire of the module.

The current conceptualization of the metabolism of acetaldehyde produced from ethanol is that it is rapidly metabolized by the protein product of *Aldh2* (aldehyde dehydrogenase 2 family), which resides in the mitochondria [312]. There is, however, ample evidence for the involvement of *Aldh1a1* when higher levels of acetaldehyde are present. Thus, the contribution

of *Aldh1a1* to acetate AUC may be primarily evident under conditions of high ethanol clearance rates.

**Insights from acetate analysis**

Inclusion of acetate in this study was done in a hypothesis generating manner. Our results implicating expression of *Adh1* and *Adh4* as major players influencing acetate exposure agree with the genetic studies on AUD that have consistently found an association between AUD and the alcohol metabolism genes and, furthermore, support the importance of gene regulation, including gene expression, that is believed to modulate many complex traits. Could the genetic link between *ADH* genes, namely the *ADH1* and *ADH4* genes, and AUD be due to acetate and, if so, what is the mechanistic explanation?

To explain – at the molecular level – the overwhelming evidence linking alcohol metabolizing genes to AUD, the prevailing hypotheses have predominantly focused on acetaldehyde. Acetaldehyde is a toxic intermediate whose accumulation leads to adverse reactions such as flushing, nausea, and rapid heartbeat (i.e., tachycardia). As a result, scientists have postulated that genetic polymorphisms leading to elevated acetaldehyde levels may protect against AUD. Below we will discuss the possible reinforcing effects of acetaldehyde, followed by a motivating hypothesis that acetate may contribute – as least in part – to genetic link between the alcohol metabolizing genes and AUD. Importantly, we note that this hypothesis is not mutually exclusive to a role of acetaldehyde.

**Acetaldehyde as the molecular explanation for the genetic association of alcohol metabolism genes with alcohol use disorder**

The strong protective effects of *ALDH2*2* on AUD [313-315] serves as the paradigm for the "aversion" hypothesis; the polymorphism renders the protein product unable to catalyze

acetaldehyde conversion to acetate. Indeed, both *in vivo* and *in vitro* studies have clearly demonstrated elevated acetaldehyde levels for those possessing the *ALDH2*2* allele [52, 316]. Consequently, individuals with *ALDH2*2* are "protected" against the development of AUD [317, 318]. SNPs causing other, less dramatic changes to alcohol metabolism kinetics are likewise explained by way of at least transiently increasing acetaldehyde accumulation; however, the causal relationship between the SNPs and acetaldehyde buildup are inconsistent. For example, *ADH1B*2*, *ADH1B*3*, and *ADH1C*1* have not been associated with elevations in acetaldehyde [319], and the physiological reactions thought to deter drinking are not readily observable. Moreover, if acetaldehyde protects against AUD by making consumption unpleasant, it raises the question as to if and how the corresponding genetic polymorphisms in *ADH* genes that modulate acetaldehyde levels contribute to the motivational and rewarding effects that lead to the development of AUD.

To reconcile this issue, research has turned to the potential role of acetaldehyde in the brain. The mechanism by which acetaldehyde reaches the brain after ingestion is controversial. Two general possibilities have been proposed: systemic circulation of acetaldehyde after metabolism of ethanol in liver (peripheral) or local metabolism of ethanol to acetaldehyde in the brain (central). Several lines of evidence suggest peripherally generated acetaldehyde does not reach the brain in appreciable amounts: 1) the blood brain barrier likely metabolizes most acetaldehyde to acetate before it can enter the brain [320-322] and 2) liver ALDH rapidly converts acetaldehyde to acetate so that very low levels of acetaldehyde are detected in blood [323]. In contrast, ethanol is able to easily cross the blood brain barrier [324] and has been shown to be metabolized to acetaldehyde in rodent brains [325-327]. Unlike the liver, its oxidation to acetaldehyde here is mostly attributable to the catalase pathway [326, 328-330] and,

to a lesser extent, the cytochrome P450 2E1 pathway [330]. This is because catalase is expressed highly in the brain [331], whereas there is a general lack of class I ADH [332]. It, however, needs to be remembered that catalase requires hydrogen peroxide ($H_2O_2$) to convert ethanol to acetaldehyde and brain levels of $H_2O_2$ are necessarily maintained at low levels.

Numerous studies have proposed – indirectly and speculatively – that acetaldehyde can mediate the pharmacological effects of ethanol leading to addiction. The history of such studies begins with the studies of Amit et al. who administered acetaldehyde directly into brain and a long history of studies trying to link excessive ethanol consumption to formation of condensation products of acetaldehyde and transmitter mono-amines to form tetrahydropalmatine (THP) like products in brain. The theories emanating from such studies have been rejected because of difficulties in finding acetaldehyde in brain at levels expected by the studies of Amit and difficulties in demonstrating THP and other like alkaloids in brain of animals given ethanol and blood or brain humans who consume ethanol. The current common approach is to study enzymatic manipulations that alter the production and degradation of acetaldehyde. Results have indicated that both peripheral and central acetaldehyde may contribute to drinking behavior. For instance, inhibition of ADH1 reduced ethanol-induced conditioned place preference (CPP), suggesting that peripherally derived acetaldehyde contributes to the motivational effect of alcohol [333] Supporting central formation, acquisition of CPP was blocked by inhibiting brain catalase activity [334]. Furthermore, peripheral and central sequestration of acetaldehyde – using L-cysteine or D-penicillamine – reduced both acquisition and maintenance of oral ethanol self-administration behavior [335-338]. Finally, peripheral [333, 339-341] and central [342, 343] administration of acetaldehyde mimics alcohol's neurobehavioral effects. But while the pharmacological effects are clear, the role of acetaldehyde is less so. The aforementioned studies

may intimate a role for acetaldehyde, but none rule out the possibility that ethanol exerts its effects through its further downstream metabolite, acetate.

**An alternative hypothesis: acetate as the molecular explanation for the genetic association of alcohol metabolism genes with alcohol use disorder**

Here we offer an alternative hypothesis to coalesce the genetic and behavioral studies. Namely, we suggest acetate as a key contributor to the genetic component of AUD, and this perspective simplifies and unifies many of the observations to date. From a behavioral perspective, most studies implicating acetaldehyde do not eliminate a role for acetate, because experimental designs used to manipulate acetaldehyde production and accumulation have the same influence on acetate. In addition, the peripheral effects of ethanol are more easily explained by acetate; unlike acetaldehyde, there is little debate that higher acetate levels are detected in blood after ethanol intake [344-346] and significant amounts reach the brain [347]. Likewise, the metabolic barrier presented by ALDH across the blood brain barrier produces additional acetate that can access the brain. Ethanol escaping metabolism in liver and reaching the brain may also be oxidized, thereby making it possible that the central effects of ethanol/acetaldehyde are mediated by acetate. The possibility that brain tissue metabolizes ethanol still needs to be demonstrated under measured alcohol drinking situations.

From a genetic perspective, a role of acetate as a contributor to AUD is also a reasonable hypothesis. Peripherally derived acetate reaches the brain in large amounts, explaining the association between alcohol metabolizing genes and AUD that has been rigorously determined through genetic studies. In contrast, if catalase is indeed the main ethanol oxidation machinery in brain, and if acetaldehyde indeed does not reach brain in appreciable amounts, it seems unlikely

that acetaldehyde could be the main mediator of AUD because *ADH* and *ALDH* genes are the ones associated with AUD.

**Acetate from alcohol exerts its effects through histone acetylation**

One possible mechanistic explanation of how acetate influences AUD at the molecular level is through histone acetylation. Previous work showed that orally delivered acetate supplementation in rats increased brain histone acetylation state. Specifically, acetate increased the acetylation of brain histones while having no effect on liver histones. Additionally, this phenomenon was mediated through inhibition of histone deacetylase (HDAC) protein levels – and in particular HDAC2 – without altering histone acetyltransferase (HAT) activity [348]. Separate studies confirmed that HDAC2 protein levels are decreased in the amygdala of alcohol preferring P rats after acute alcohol exposure and also showed that P rats innately express greater HDAC2 protein in the amygdala than non-preferring NP rats [349, 350]. Besides displaying higher alcohol preference compared to NP rats,[111, 351] P rats also exhibit heightened anxiety-like behavior [352, 353]. Voluntary ethanol consumption (7% or 9% ethanol) produced anxiolytic effects only in P rats, and HDAC inhibitors and HDAC2 siRNA normalized the anxiety like behavior and attenuated the excessive alcohol intake [349, 350]. Taken together, these results provide circumstantial evidence that acetate could mediate some of the behavioral properties of aspects of alcohol and alcohol consumption.

Mews et al. [354] established with greater certainty the fact that acetate derived from alcohol is responsible for altering brain histone acetylation. Using *in vivo* stable isotope labeling in mice, the authors established that 1) acetate from alcohol metabolism contributed to rapid acetylation of histones in brain both in the hippocampus and prefrontal cortex, 2) this occurred in a chromatin-bound acetyl-CoA synthetase 2 (ACSS2) dependent fashion, and 3) much of this

acetate was produced peripherally rather than centrally. Follow up functional experiments showed that a myriad of genes undergo hyperacetylation in the dorsal hippocampus and, subsequently, expression changes, in response to alcohol treatment. To ascertain that acetate *per se* was responsible for the observed acetylation and expression changes in the brain, and not another effect resulting from alcohol metabolism, they next supplemented primary hippocampal neurons with acetate *ex vivo*. Many of the genes upregulated *in vivo* and *ex vivo* overlapped. Also noteworthy, a separate *in vivo* study found alcohol induces altered expression for similar genes in hippocampal neurons, suggesting acetate is the mediator of this effect [355]. In general, genes whose expression was modulated displayed enrichment in functions that could be related to the development of AUD (e.g., neuronal plasticity), and several individual genes have been directly linked to alcohol use, drug related behavior, and addiction. As final evidence that alcohol's rewarding effects may also be mediated in part through acetate, the authors compared CPP in wild type vs ACSS2-knockdown mice where the knockdown removes the pathway by which alcohol derived acetate promotes acetylation and gene expression effects. Indeed, knockdown of ACSS2 in the dorsal hippocampus greatly reduced the expression of CPP.

**Acetate from alcohol exerts its effects through brain metabolism for energy.**

These epigenetic changes may explain some of the persistent aspects of AUD attributable to acetate derived from alcohol. Another, more immediate, potential avenue in which acetate mediates some of the cerebral effects of alcohol is through caloric "reward" in the brain. Acetate is endogenously produced by the body even in the absence of ethanol. Thus, in essence, acetate from alcohol can enter the same metabolic paths as acetate not generated from alcohol. While the brain normally relies on glucose for energy, acute alcohol exposure reduces glucose metabolism and increases acetate uptake [356, 357]. Changes in brain glucose metabolism are sensitive to

past histories of alcohol use [358]. For instance, circulating acetate levels [359] and brain acetate uptake [356, 360] are increased in heavy drinkers compared to their social drinking counterparts. Furthermore, those reporting a greater history of alcohol consumption likewise had greater acetate uptake [356]. Enculescu et al. [361] observed changes in protein abundances for key metabolic enzymes (e.g., glycolysis, trafficking, the cytoskeleton, and excitotoxicity) in AUD, and suggested a switch from glucose to acetate utilization in AUD brain. Researchers have therefore hypothesized that cessation of drinking in those afflicted with AUD may lead to discomfort due to loss of acetate availability for brain metabolism, thereby promoting continued drinking [356, 360]. Calories from alcoholic beverages in heavy drinkers can be as much as 50% of total energy intake.[362] Notably, acetate supplementation in rats – made physically dependent on alcohol – decreased alcohol withdrawal tremors [363]. After completion of this work, we next wanted to probe the role APA plays in alcohol related phenotypes using a systems genetics approach, which requires their identification and quantitation in the transcriptome.

# CHAPTER III

# APTARDI PREDICTS POLYADENYLATION SITES IN SAMPLE-SPECIFIC TRANSCRIPTOMES USING HIGH THROUGHPUT RNA SEQUENCING AND DNA SEQUENCE[2]

[2]This chapter was previously published in: Lusk R., et al., Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. *Nat Commun*. 12, 1652 (2021).

## Introduction

Alternative polyadenylation (APA) is gene regulation mechanism by which a single gene encodes multiple RNA isoforms with different polyadenylation (polyA) sites [104] (i.e., different transcription stop sites/3' termini). Most APA sites lead to identical protein products but variable 3' UTR lengths [102]. APA has been associated with disease through many transcripts displaying APA (e.g. cardiac hypertrophy [364], oculopharyngeal muscular dystrophy[365, 366], breast cancer, and lung cancer [367]) and APA in an individual transcript (e.g. Fabry disease [368], amyotrophic lateral sclerosis [369], metachromatic leukodystrophy [370], and facioscapulohumeral muscular dystrophy [371]). Furthermore, differences in expression of APA transcripts have been implicated in diseases [169] and are recognized as risk factors in complex diseases [170]. Indeed, research suggests individual susceptibility to complex diseases is mainly due to variation in gene regulation processes – such as APA – rather than variation in protein coding sequence [96-99]. APA's impact is expected given that it is pervasive, with more than 70% of human genes subjected to APA [163], and also far-reaching, as it modulates mRNA stability, translation, nuclear export, and cellular localization, as well as the localization of the encoded protein [102, 103] – often times through differences in miRNA binding availability.

APA patterns are tissue specific [184, 185], and "choice" of polyA sites can be influenced by physiological, environment, and disease states [104, 183]. This dynamic may explain – at least in part – why polyA sites are often under annotated [191] and, furthermore, why (the often times sparse) prior annotation is typically not relevant to the given set of experimental conditions [372]. As a result, polyA sites often need to be re-defined for the sample(s) of interest to gain insight into the role of APA in various processes and diseases (e.g. are certain APA transcripts biomarkers of, or therapeutic targets for, a given disease state?). There are three broad sequencing technologies utilized to identify polyA sites: 1) short-read RNA-Seq, 2) direct 3' end RNA sequencing, and 3) DNA sequence, but each possesses inherent limitations for sample-specific identification of polyA sites (see Chapter I for a review).

To overcome current limitations, we introduce aptardi (alternative polyadenylation transcriptome analysis from RNA-Seq data and DNA sequence information). Aptardi leverages the information afforded by DNA nucleotide sequence information (from the appropriate reference genome) and RNA-Seq, as well as the predilection of transcriptome assemblers to accurately characterize splice junctions, in a use-all-data, multi-omics approach to create a modified, sample-specific transcriptome that includes information on expressed polyA sites (Figure 3.1). Specifically, harnessing the power of (supervised) machine learning, we trained aptardi to detect polyA sites from DNA nucleotide sequence and RNA-Seq read coverage by training on polyA sites identified by 3' sequencing. Using what it learned, aptardi makes predictions from DNA sequence and RNA-Seq alone, alleviating the burden of generating 3' sequencing data. The program evaluates initial transcripts in the input original transcriptome to identify expressed polyA sites in the biological sample and refines transcript 3' ends accordingly and outputs its results to a modified transcriptome (as a General Feature Format [GTF] file).

Additionally, aptardi's input is simple to compile and its output is easily amenable to downstream analyses such as quantitation and differential expression.

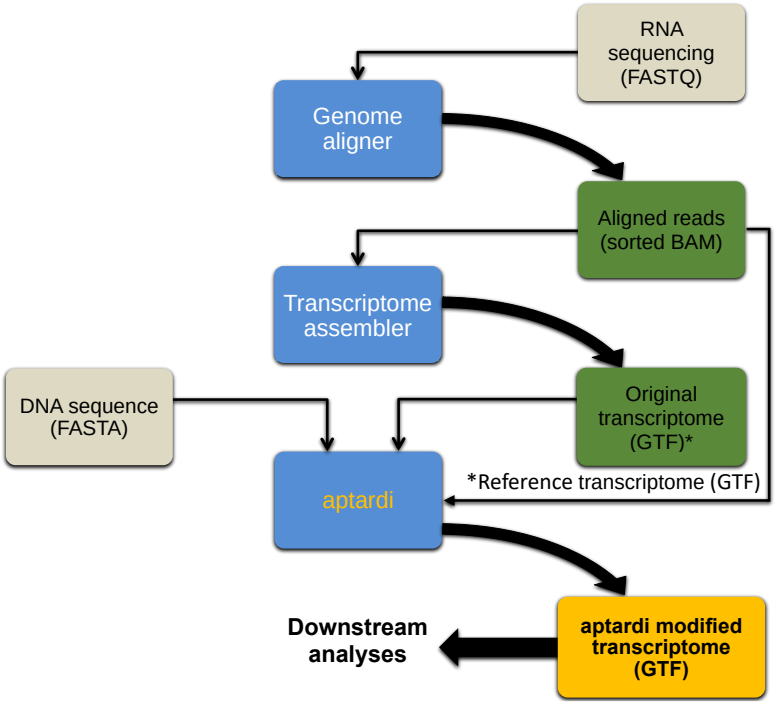



**Figure 3.1. Overview for using aptardi.** Aptardi requires three files as input: 1) FASTA file of DNA sequence with headers by chromosome, 2) sorted Binary Alignment Map (BAM) file of reads aligned to the genome, and 3) General Feature Format (GTF) file of transcript structures. Blue boxes represent software. Yellow writing/boxes indicate aptardi incorporation. Note transcript structures can be derived from a reference transcriptome (i.e., Ensembl annotation) in lieu of the original transcriptome generated from a transcriptome assembler.

## Methods

### Aptardi design

The overall goal of aptardi is to accurately identify the polyA sites of expressed transcripts in a given biological sample. Specifically, aptardi analyzes the modified 3' terminal exon (see the Transcript processing section below for details on 3' terminal exon modification) of previously annotated transcripts and, using relevant RNA-Seq data and DNA sequence in a machine learning environment, identifies locations of expressed polyA sites in the region. Aptardi then annotates the 3' termini to match these locations and outputs these transcript

structures to the transcriptome (in GTF format) that can be easily incorporated into to downstream analyses. Note that aptardi does not evaluate the intron chain structure of transcripts, i.e., it only examines the modified 3' terminal exon of each transcript structure and alters the 3' terminus location(s) accordingly. Also note that aptardi outputs all original transcript structures from the original transcriptome in addition to transcripts identified through its analysis, i.e., the program only adds transcripts.

**Datasets**

A total of five unique datasets, hereafter referred to as HBR, 2nd HBR, UHR, BNLx, and SHR, were subjected to aptardi's machine learning pipeline. In addition to RNA-Seq measurements, each dataset required DNA sequence, a transcriptome, and – since each was used to build a machine learning model – a "gold standard" data source providing locations of expressed, i.e. "true" polyA sites. HBR, 2nd HBR, and UHR are well-established RNA reference samples from the MAQC/SEQC consortium [373] (see RNA sequencing datasets for more details). BNLx and SHR represent two inbred rat strains: the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub) and the spontaneous hypertensive rat strain (SHR/OlaIpcv), respectively.

**DNA sequence datasets**

For BNLx and SHR, strain-specific genomes were generated from the rn6/Rnor_6.0 version of the rat genome [133] and are publicly available on the PhenoGen website. The human reference genome (hg38/GRCh38), accessed via the UCSC Genome Browser [374], was utilized for the HBR, 2nd HBR, and UHR DNA sequence datasets.

**RNA sequencing datasets**

The HBR and 2nd HBR RNA-Seq datasets were derived from the Human Brain Reference (multiple brain regions of 12 donors, Ambion, p/n AM6050), and the UHR RNA-Seq dataset was derived from the Universal Human Reference (10 pooled cancer lines, Stratagene, p/n 740000). Each of these datasets were accessed from the Sequence Read Archive (SRA) using the SRA Toolkit (v.2.8.2) as publicly available data (HBR [375]: Accession: PRJNA510978, SRA runs: SRR8360036-37; 2nd HBR and UHR [376]: Accession: PRJNA362835, 2nd HBR SRA runs: SRR5236425-30, UHR SRA runs: SRR5236455-60; BNLx and SHR [132]: Accession: GSE166117, BNLx: GSM5061950-52, SHR: GSM5061947-49). Briefly, all libraries were generated with the TruSeq stranded (HBR, 2nd HBR, UHR) or unstranded (BNLx and SHR) mRNA sample preparation kit (Illumina), sequenced on a HiSeq2500 Instrument (Illumina), and sequencing results processed to FASTQ files. The HBR RNA-Seq dataset originated from 1 µg RNA starting material, while 100 ng input was used for 2nd HBR and UHR RNA-Seq datasets. For more detailed descriptions on these publicly available data, see Palomares et al. [375] (HBR) and Schuierer et al. [376] (2nd HBR and UHR). For BNLx and SHR, RNA-seq libraries prepared from the polyA+ fraction were constructed using the Illumina TruSeq RNA Sample Preparation kit from one µg of brain RNA in accordance with the manufacturer's instructions. Four µL of a 1:100 dilution of either ERCC Spike-In Mix 1 or Mix 2 (ThermoFisher Scientific) were added to each extracted RNA sample. An Agilent Technologies Bioanalyzer 2100 (Agilent Technologies) was utilized to assess sequencing library quality. RNA samples from three biological replicates per strain were processed and sequenced [132]. All reads were paired-end but differed in read length (HBR, BNLx, and SHR: 2X100, 2nd HBR and UHR: 2X75). Individual FASTQ files were assessed for quality using FastQC (v.0.11.4,

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and, if necessary, reads were trimmed with cutadapt [377] (v.1.9.1). For the purpose of read coverage used by the aptardi algorithm, reads from technical replicates (HBR = 2, 2nd HBR = 3, UHR = 3) or biological replicates (BNLx = 3, SHR = 3) were concatenated and aligned to their respective genomes (see DNA sequence datasets section for more details) using HISAT2 [378] (v.2.1.0) with the --rna-strandness (when appropriate) and --dta options specified as recommended for transcriptome assembly with StringTie [191] (see Transcriptome datasets below for more details) and otherwise default arguments (see Supplementary Table 5 for alignment results). After alignment, SAMtools [379] (v.1.9) was used to remove unmapped reads and convert the output to a sorted Binary Alignment Map (BAM) file required as input by aptardi.

**True polyadenylation sites datasets**

True polyA sites (i.e., labels for machine learning) were taken from Derti et al. [208] for all datasets. Namely, total RNA from the same UHR and HBR RNA reference samples, as well as brain total RNA from the Sprague Dawley rat (Zyagen, p/n RR-201), were subjected to PolyA-Seq analysis to identify the genomic locations of expressed polyA sites in each sample (for more information, see Derti et al. [208]). High quality filtered polyA sites from each RNA sample were accessed using the UCSC Table Browser [380], and liftOver [381] (from the UCSC Genome Browser Group) was used to convert the genomic coordinates to the most recent human genome assembly (hg38/GRCh38) for the HBR and UHR samples, or rat genome assembly (rn6/Rnor_6.0) for the rat brain sample. PolyA sites identified in the HBR and UHR RNA reference samples were used for the corresponding HBR, 2nd HBR and UHR datasets, and those identified in the rat brain were used for both BNLx and SHR. Derti et al. [208] uploaded technical replicates to UCSC Table Browser for each RNA sample; however, since polyA sites

within 30 bases were clustered into the single site with greatest expression, and since this was done separately for each dataset, we utilized only a single dataset for each sample, i.e. technical replicates were not combined.

**Original transcriptome generation**

StringTie [191] (v.1.3.5) was used to reconstruct the transcriptome expressed in each dataset from their RNA-Seq data, hereafter referred to as the original transcriptome. Ensembl [382] (v.99) reference annotation from the respective species was provided to guide the StringTie reconstruction. We note that a user can simply use reference annotation directly, i.e. Ensembl annotation, in lieu of performing transcriptome assembly that takes into account expression, i.e. StringTie. If the RNA-Seq data were stranded, the read orientation was specified as an argument to StringTie. Transcript structures from scaffold chromosomes and unstranded contigs, if present, were removed.

**The data processing pipeline**

*Transcript processing*

Using the original transcriptome, the 3' terminal exons of transcripts were isolated. Each transcript's 3' terminal exon was extended 10,000 bases plus two times the bin size (i.e., 10,200 bases for the default 100 base bin size) similar to what has been done previously [205, 206]. Extensions overlapping any neighboring transcripts (on the same strand) were shortened to remove the overlap. RNA-Seq coverage at single-nucleotide resolution was obtained via bedtools genomecov (BEDtools [383]; v2.29.2) and, similar to the criteria employed by Ye et al. [206] and Miura et al. [384], used to refine each of these 3' terminal exons, hereafter referred to as modified 3' terminal exons (see Supplementary Information for more details.) The refinement

step either shortened the extended 3' terminal exon or kept it the same length to give the modified 3' terminal exon.

*Feature extraction*

Features were engineered in 100 base increments along the modified 3' terminal exon, referred to hereafter as bins. For each bin, a total of 27 features were engineered and can be broadly classified as being derived from DNA sequence or RNA-Seq data. In both cases, information from the local environment, i.e., the 100 bases upstream and downstream the bin, as well as the bin itself, (300 bases total) was used.

*DNA sequence features*

The choice of DNA sequence features was made through a combination of an exhaustive literature review [178, 187, 385-392] and evaluation of other algorithms that use DNA sequence to predict polyA sites [180, 181, 393, 394]. Perhaps the most well-known indicator of polyadenylation is the polyadenylation signal (PAS), a conserved hexamer located ~10-35 nucleotides upstream the polyA site. Overrepresented sequences of DNA, or DNA sequence elements, also influence polyadenylation, and the location of these sequences are often described relative to the PAS. As such, DNA sequence features were engineered by first identifying the presence of several known PAS's. Specifically, for each bin, a six base sliding window scanned a predefined region to detect the presence or absence (binary indicator of 1/-1) of 1) the canonical PAS (AATAAA), 2) its major variant (ATTAAA), 3) a second common variant (AGTAAA), and 4) any one of nine other minor variants (AAGAAA, AAAAAG, AATACA, TATAAA, GATAAA, AATATA, CATAAA, AATAGA[178, 187, 385, 386]) for four total PAS features. Subsequently, regions relative to the PAS (if present, otherwise predefined regions relative to the current bin) were likewise scanned using a sliding window approach to determine frequency of

the following known DNA sequence elements: 1) a G-rich region downstream the PAS, 2) a downstream region near the PAS enriched in TTT, 3) a downstream region near the PAS enriched in GT/TG, and 4) a downstream region near the PAS enriched in GTGT/TGTG, 5) a T-rich region immediately downstream of the PAS, 6) a T-rich region upstream the PAS, 7) a TGTA/TATA-rich region upstream the PAS, and 8) a AT-rich region upstream and downstream the PAS[178, 187, 387-392] for an additional 8 features (12 DNA sequence features total). If the frequency of the given DNA sequence element was above an enrichment threshold, the feature was encoded 1, otherwise -1. (See Supplementary Information for more details.)

*RNA sequencing features*

From the RNA-Seq data, coverage at single nucleotide resolution was determined using BEDtools[383] (v2.29.2). The approach for designing RNA-Seq features was to exploit localized fluctuations in RNA read coverage similar to that implemented by tools designed for APA-specific analysis from RNA-Seq data [205-207]. Intuitively, upstream but in close proximity to the end of a transcript, coverage is expected to begin to decrease gradually until its end. As a result, changes in expression were utilized when designing RNA-Seq features in two scenarios: 1) intra- and 2) inter-bin. In both cases, three regions were defined: an upstream region, a middle region, and a downstream region (Supplementary Figure 4). Changes in expression between these regions were quantified using various mathematical combinations of coverage values in each region to generate 14 unique features (see Supplementary Information for more details). To account for local variability in RNA-Seq coverage, median coverage values in each region were used.

A final feature was derived from the original transcriptome. If the 3’ base of any annotated transcript from the original reconstruction was located within a bin, this feature was

encoded 1, otherwise -1. Supplementary Figure 5 summarizes the data processing pipeline prior to machine learning.

**Building aptardi**

The machine learning task is two class classification (polyA site or no polyA site) of each 100 base bin. Supervised learning was used where labels for training were provided from the polyA sites datasets (see PolyA sites datasets for more details). A BiLSTM [220, 221, 395, 396] was implemented using the Keras (v.2.3.1) wrapper for TensorFlow (v.2.0.0). This machine learning paradigm was chosen because of its design to analyze sequential data, i.e. it takes into account all the 100 base bins of a given transcript when learning model parameters for each individual bin. Each direction of the biLSTM consisted of 20 nodes (40 total), and this layer was followed by a fully connected dense layer with a sigmoid activation function that outputs a probability value. Of note, the biLSTM outperformed traditional classifiers such as Random Forest (RF) and Support Vector Machine (SVM) models (biLSTM: AP = 0.58, F-measure = 0.52; RF: AP = 0.42, F-measure = 0.39; SVM: AP = 0.37, F-measure = 0.33; numbers shown are on the testing set using the HBR dataset).

**Training aptardi**

To prevent duplicate bins, overlapping modified 3' terminal exons were merged prior to training. Additionally, all merged modified 3' terminal exons were masked to a length of 300 bins (30,000 bases total) to generate equal lengths, which is required for the sequential model. A total of 778,166 bins were present, of which 42,977 possessed a polyA site out of 94,322 polyA sites annotated by the PolyA-Seq data (for the HBR dataset). Merged modified 3' terminal exons were split into 60/20/20 training, validation, testing sets, respectively. Quantitative measures were standardized using the training set, and the training set was used to build the model in 25

epochs. Due to the high imbalance of the data, class weights were used during training. Model weights were optimized using a binary cross entropy loss function and Adam [397] optimizer. Precision and recall metrics on the training and validation sets were monitored during training to prevent overfitting, and the model that produced the minimum loss was kept. For evaluation purposes (see Results), individual prediction models were generated from each of the five datasets.

**Evaluating aptardi**

Precision, recall, and F-measure at the default probability threshold (0.5) were used to evaluate model performance defined as follows:

$$P = \frac{T_p}{T_p + F_p} \tag{3.1}$$

$$R = \frac{T_p}{T_p + F_n} \tag{3.2}$$

$$F = 2 * \frac{(P * R)}{(P + R)} \tag{3.3}$$

where $T_p = true\ positive$, and $F_p = false\ positive$, $F_n = false\ negative$, $P = precision$, $R = recall$, and $F = F - measure$

To generalize model performance over the range of probability thresholds, average precision was used in place of the receiver operating curve due to the highly imbalanced nature of the data [398] (far fewer bins with polyA sites than bins without a polyA site):

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{3.4}$$

where $AP = average\ precision$ and $R_n$ and $P_n$ are the precision and recall at the $nth$ threshold, respectively.

**Integrating aptardi results with the original transcriptome**

For 100 base bins where a polyA site is predicted, transcript structures are annotated to the 3' most base position unless either 1) the input transcript's stop site is already in the region or 2) the 3' most base position is within 100 bases of the input transcript's stop site. Any aptardi transcript structures were added to the original transcriptome, and this aptardi modified transcriptome was outputted as a GTF file.

**Choice of bin size**

To assess the impact of bin size, 25, 50, and 150 base bins were used to train the aptardi prediction model on the HBR dataset in an otherwise identical manner to the original 100 base bin. The average precision on the testing set was 0.41, 0.51, and 0.61 for the 25, 50 and 150 base bins, respectively, compared to 0.58 for the original 100 base bin. Additionally, the F-measure on the testing set was 0.35, 0.46, and 0.52, respectively, compared to 0.52 for the for the original 100 base bin. Based on these results, 100 base bins were utilized to build the pre-existing aptardi prediction model. Higher resolution bin sizes may be appropriate depending on the dataset (e.g., RNA-Seq library preparation), species, etc. and bin size can be specified as a parameter when using the aptardi pipeline.

**Software**

A user has the option of using the pre-existing aptardi prediction model or building a prediction model (if a reliable true polyA sites dataset is available). The pre-built model, i.e., the aptardi prediction model, provided on the aptardi GitHub repository (https://github.com/luskry/aptardi) was generated from the HBR dataset [375]. Several other algorithm options are available.

**TAPAS analysis**

TAPAS (Tool for Alternative Polyadenylation site AnalysiS) predicts the locations of polyA sites from RNA-Seq and reference annotation (genome or transcriptome) [207]. Its performance was evaluated on HBR, specifically using the HBR RNA-Seq data and StringTie/Ensembl original transcriptome. Since TAPAS makes predictions on noncoding sequence coordinates of transcript models (i.e., 3' UTRs), the 3' terminal exon of each transcript was provided as the noncoding region, and default arguments were used. The TAPAS polyA site prediction program (APA_sites_detection) was used here and accessed via its GitHub repository (https://github.com/arefeen/TAPAS).

**APARENT analysis**

APARENT (APA REgression NeT) predicts the locations of polyA sites from DNA sequence [399]. Its performance was evaluated using the human reference genome (hg38/GRCh38) and transcript structures generated from the HBR StringTie/Ensembl original transcriptome. Specifically, similar to that for aptardi, the 3' terminal exon of each of the 113,923 transcript models in the original transcriptome were extracted. Since APARENT requires DNA sequences to be greater than or equal to 205 nucleotides and less than or equal to 10,000 nucleotides, 20,658 3' terminal exons were removed from the analysis. For the remaining 93,265 3' terminal exons, the DNA sequence was extracted using BEDtools (v2.29.2). For 3' terminal exons on the negative strand, the reverse complement sequence was used. The APARENT model (aparent_large_lessdropout_all_libs_no_sampleweights.h5) from its GitHub repository (https://github.com/johli/aparent) was used to predict the locations of polyA sites using default parameters.

**CFIm25 knockdown analysis**

RNA-Seq from HeLa cells and RNA-Seq after RNA interference on HeLa cells was used to generate the control RNA-Seq dataset and the treatment CFIm25 knockdown RNA-Seq dataset that induces APA switching, respectively. The RNA-Seq datasets were accessed from SRA as publicly available data (Accession: PRJNA182153, control SRA run: SRR1238549, CFIm25 knockdown SRA run: SRR1238551). The RNA-Seq library preparation was unstranded, and 100 base paired end reads were sequenced on an Illumina HiSeq 2000 instrument (see Masamha et al. [400] for more details). Reads were processed in a manner identical to all other datasets to produce a sorted BAM file (see Supplementary Table 6 for alignment results). An original transcriptome using StringTie/Ensembl (without the read orientation argument) and an aptardi modified transcriptome were generated for each RNA-Seq dataset (four total). The aptardi prediction model produced using the HBR dataset was used to generate the aptardi modified transcriptomes.

**Mouse tissue analysis**

RNA-Seq from mouse brain and liver tissues were taken from Li et al. [401] and accessed from SRA as publicly available data (Accession: PRJNA375882, brain SRA runs: SRR5273637 and SRR5273673, liver SRA runs: SRR5273636 and SRR5273672). The two SRA runs per tissue represent technical replicates from one biological sample – namely six-week old female C57BL/6JJcl mice. Briefly, libraries were constructed using polyA selection and the Illumina TruSeq RNA-Seq library protocol and sequenced using an Illumina HiSeq platform to generate 100 base, paired end, and unstranded reads. Reads were trimmed for quality and adapter content with Trimmomatic [402] (v.0.39), and the two technical replicates per tissue were aligned together to the mm10/GRCm38 genome using HISAT2 (v.2.1.0; see Supplementary Table 3 for

alignment results). An original transcriptome was generated using StringTie (v.1.3.5) and the mouse Ensembl (v.102) annotation as a guide and otherwise default settings. BALB/c mouse brain and liver PolyA-Seq data from Derti et al. were used to define the tissue-specific true polyA sites. Namely, high quality filtered true polyA sites were accessed using the UCSC Table Browser, and liftOver (from the UCSC Genome Browser Group) was used to convert the genomic coordinates to the most recent mouse assembly. Tissue specific polyA sites in brain and liver were defined as polyA sites in the given tissue PolyA-Seq dataset not within 100 bases of any polyA site in the other tissue PolyA-Seq dataset. The mouse reference genome (mm10/GRCm38), accessed via the UCSC Genome Browser was used for DNA sequence for both tissues. Comparisons between the polyA sites from the given PolyA-Seq data and transcriptome were done by defining that a transcript annotated a polyA site if its 3' terminus was within 100 bases of the site.

**Rat differential expression analysis**

To generate a single transcriptome representing both rat strains, their genome-aligned RNA-Seq data were merged using SAMtools [379] (v.1.9) followed by the production of an original transcriptome using the merged RNA-Seq dataset and StringTie/Ensembl (without the read orientation argument). The aptardi modified transcriptome was produced using the original transcriptome as input, along with the merged RNA-Seq data, the rn6/Rnor_6 DNA sequence (accessed via the UCSC Genome Browser [374]), and the aptardi prediction model built from HBR. RSEM[257] (v.1.2.31) was used to estimate the abundances of the isoforms identified within each transcriptome (the original transcriptome and aptardi modified transcriptome). Prior to quantitation, transcripts from scaffold chromosomes and unstranded contigs were removed from both transcriptomes. Isoform level expression estimates were determined for each

biological sample (BNLx = 3, SHR = 3). Isoforms without at least 50 counts in two of the three biological replicates for at least one strain were removed, and differential expression between the two strains (with BNLx as reference) were evaluated using DESeq2 [264] (v.1.28.0) for the remaining set of isoforms in each transcriptome (Supplementary Figure 6 summarizes these analysis steps). A significance threshold of 0.001 was applied to the unadjusted p-values to allow for comparisons across the two datasets (original transcriptome and aptardi modified transcriptome) that differ in the number of transcripts tested.

**Data availability**

All data are publicly available. The genomic sequence data that support the findings of this study are available on the UCSC Genome Browser (human: hg38/GRCh38, rat: rn6/Rnor_6.0, mouse: mm10/GRCm38, http://genome.ucsc.edu/) and PhenoGen (BNLx: BN-Lx/CubPrin, SHR: SHR/OlaIpcvPrin, https://phenogen.org/). The polyadenylation sites from PolyA-Seq data that support the findings of this study are available on the UCSC Table Browser (https://genome.ucsc.edu/). The polyadenylation sites from PolyA_DB and PolyASite 2.0 that support the findings of this study are available on their respective websites (PolyA_DB: https://exon.apps.wistar.org/PolyA_DB/v3/, PolyASite 2.0: https://polyasite.unibas.ch/). The RNA sequencing data that support the finding of this study are available on the NCBI Sequence Read Archive (Human Brain Reference RNA sequencing: Accession: PRJNA510978 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA510978/], SRA runs: SRR5236425-30; 2nd Human Brain Reference and Universal Human Reference RNA sequencing: Accession: PRJNA362835 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA362835], 2$^{nd}$ HBR SRA runs: SRR5236425-30, UHR SRA runs: SRR5236455-60; Control vs CFIm25 knockdown RNA sequencing: Accession: PRJNA182153

[https://www.ncbi.nlm.nih.gov/bioproject/PRJNA182153/], control SRA run: SRR1238549, CFIm25 knockdown SRA run: SRR1238551; Mouse tissue analysis RNA sequencing: Accession: PRJNA375882[https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA375882 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA375882], brain SRA runs: SRR5273637 and SRR5273673, liver SRA runs: SRR5273636 and SRR5273672). The BNLx and SHR RNA sequencing that support this study have been deposited in NCBI Sequence Read Archive with the primary accession code GSE166117 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE166117 ] (BNLx: GSM5061950-52; SHR: GSM5061947-49).

**Code availability**

The software aptardi [403] is maintained on its GitHub repository (https://github.com/luskry/aptardi).

## Results

### Construction of multi-omics model for identification of polyadenylation sites

The initial dataset used for developing the aptardi model was derived from Human Brain Reference [373] (HBR) RNA using Illumina's TruSeq stranded mRNA sample preparation kit to generate 100 base, paired end reads [375]. The transcriptome reconstruction contained 113,923 transcripts (excluding those from scaffold chromosomes) with 94,369 unique transcript termini, and the corresponding PolyA-Seq data contained 94,322 polyA sites in this sample. Throughout this manuscript we refer to the polyA sites identified from PolyA-Seq [208] as "true" polyA sites to distinguish them from polyA sites predicted by a computational algorithm, but we acknowledge that there are false negatives and false positives among the PolyA-Seq derived polyA sites. After transcript processing (see Methods) and integration with the PolyA-Seq data –

generated from the same HBR RNA – 70,748 transcript models with zero to 50 true polyA sites per transcript model were used for learning and evaluating the aptardi prediction model. The modified 3' terminal exons of these transcript models were binned into 100 base increments for machine learning, and 14 RNA-Seq features, 12 DNA sequence-related features, and one feature derived from the original transcriptome were calculated for each bin.

As examples of DNA sequence derived features, the presence of the three strong polyA signals (5'-AATAAA-3', 5'ATTAAA-3', and 5'-AGTAAA-3') in each 100 base bin as a function of whether the bin also contained a polyA site are shown in Figure 3.2a. Each strong polyA signal demonstrated enrichment (AATAAA: $\chi^2 =$ 80,837, p-value < 0.0001; ATTAAA: $\chi^2 =$ 15,012, p-value < 0.0001; AGTAAA: $\chi^2 =$ 1,378, p-value < 0.0001 ). For instance, of the 100 base bins that possessed a true polyA site via PolyA-Seq, over half also possessed AATAAA. In contrast, only approximately 10% of bins that did not contain a polyA site had the AATAAA signal. This enrichment was observed for all binary features, i.e. all the DNA sequence features and the original transcriptome end location feature (Supplementary Figure 1a). The strong polyA signals were also independently associated with the presence of a polyA site. Likewise, the distribution of the quantitative RNA-Seq features, e.g. the inter-bin RNA-Seq features (Figure 3.2b) differed based on the presence or absence of a true polyA site (although no one feature distinguishes the true polyA sites perfectly) and this was also seen for the intra-bin RNA-Seq features (Supplementary Figure 1b) . Furthermore, features derived from DNA sequence and RNA-Seq were independent of one another across omics type but were often correlated within an omics category (Figure 3.2c).

When aptardi was built using only features derived from RNA-Seq or only features derived from DNA sequence, the average precision (AP) in the testing dataset was significantly

greater than simply relying on the polyA sites identified in the original transcriptome (Figure 3.2d). Furthermore, when the RNA-Seq features and the DNA sequence features were combined, the multi-omics model had higher AP than either single-omics model (multi-omics AP = 0.58, DNA-only AP = 0.41, RNA-only AP = 0.44). Using a specific prediction threshold (probability > 0.5), the precision in the multi-omics model (0.74) increased from the DNA-only model (0.65) but only modestly increased from the RNA-only model (0.71); however, the recall dramatically improved compared to both single-omics models (multi-omics recall = 0.39, DNA-only recall = 0.18, RNA-only = 0.24; Figure 3.2e). The F-measure was similarly greater in the multi-omics model than either single-omics model (multi-omics = 0.51, DNA-only = 0.28, RNA-only = 0.36). In addition, performance results were consistent across five random splits of the data for training/validation/testing, and the results displayed above are the averages across the five splits.

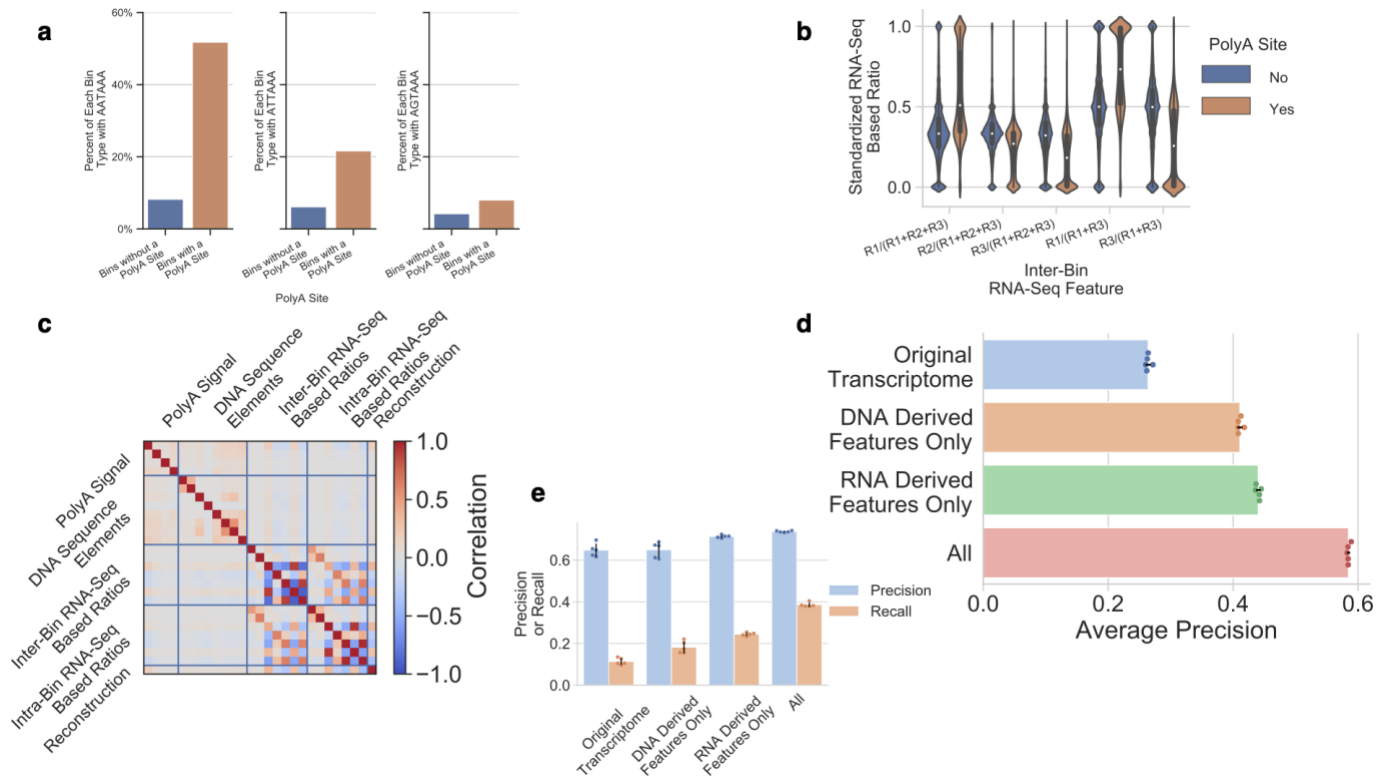



**Figure 3.2. DNA sequence and RNA sequencing (RNA-Seq) features are individually associated with polyadenylation (polyA) sites. a,** The percent of 100 base bins containing each of the three strong polyA signals stratified by the bin not containing (blue) or containing (orange) a polyA site. **b,** Distribution of the inter-bin RNA-Seq features for each 100 base bin stratified by the bin not containing (blue) or containing (orange) a polyA site (RNA-Seq ratio features were standardized using the training set). **c**, RNA-Seq features and DNA sequence features display little correlation (two-sided Pearson Product-Moment) across omics type. The combination of RNA-Seq information and DNA sequence information improves **d**, average precision and **e**, precision and recall at a specific prediction threshold (probability > 0.50) over each separately. For both d and e, data are presented as mean values +/- standard deviation on the test set (n = 5 random train-validate-test splits). Data shown are from the Human Brain Reference dataset.

The relative contributions of the individual DNA sequence features were further explored by generating aptardi models for each DNA sequence feature that either 1) included all other features but the DNA sequence feature or 2) removed all other DNA-derived features but the DNA sequence feature (i.e., all of the RNA-Seq features and the original transcriptome feature were also still included) and evaluating performance on the testing split. Unsurprisingly, the greatest reduction in performance from leaving out any single DNA sequence feature came from

removing the canonical polyA signal (AP = 0.52 vs 0.58 in full model, F-measure = 0.45 vs 0.52 in full model), and the greatest improvement by including any single DNA sequence feature was from this feature as well (AP = 0.54 vs 0.47 without any DNA-derived features , F-measure = 0.47 vs 0.39 without any DNA-derived features).

**Evaluation of the generalizability of aptardi**

To evaluate the generalizability of the aptardi prediction model, we asked two questions: 1) does the performance of the aptardi prediction model, built on the HBR dataset, remain consistent across diverse datasets, and 2) are the performances of prediction models built on alternative datasets comparable to the aptardi prediction model (built from the HBR dataset)? To answer these questions, we analyzed four alternative datasets. These datasets were chosen because they had sufficient similarities and differences to assess the applicability of the aptardi prediction model (Supplementary Table 1). Namely, an additional Human Brain Reference RNA dataset was included that was derived from the same Human Brain Reference RNA sample but processed and sequenced in another laboratory ($2^{nd}$ HBR), and this laboratory also produced another dataset we included from Universal Human Reference (UHR) RNA[376]. To include a cross-species comparison and to examine similar tissue across two genetically different individuals, we also used data derived from two inbred rat strains; the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub; BNLx) and the spontaneously hypertensive rat strain (SHR/OlaIpcv; SHR). All true polyA sites were derived from 3' sequencing PolyA-Seq data; true polyA sites for the HBR and UHR datasets were from the same corresponding RNA, whereas the true polyA sites for the two rat datasets were derived from Sprague Dawley rat brain RNA[208].

We first examined whether users can confidently apply the aptardi prediction model, built from the HBR dataset, on their own datasets (i.e., on a dataset not used to train the model) by comparing its performance on the four alternative datasets not used to train the model. The AP of the HBR-based aptardi prediction model across the four other datasets ranged from 0.55 to 0.63, whereas the AP of this HBR aptardi prediction model on its own HBR dataset was 0.65 (Figure 3.3; orange bars). Specifically, its performance on the other human RNA samples (2nd HBR and UHR) only differed in AP by two percentage points (AP = 0.63 for each), but on the BNLx and SHR rat brain datasets the HBR-based aptardi prediction model performed more modestly (AP = 0.55 for each). Similar results were observed for F-measures (HBR = 0.56, 2nd HBR = 0.51, UHR = 0.53, BNLx = 0.47, SHR = 0.48). The major differences between the two rat datasets and the HBR dataset include species, strandedness of the library preparation (rat samples were unstranded), and the inexact matching between RNA-Seq and PolyA-Seq RNA sources.

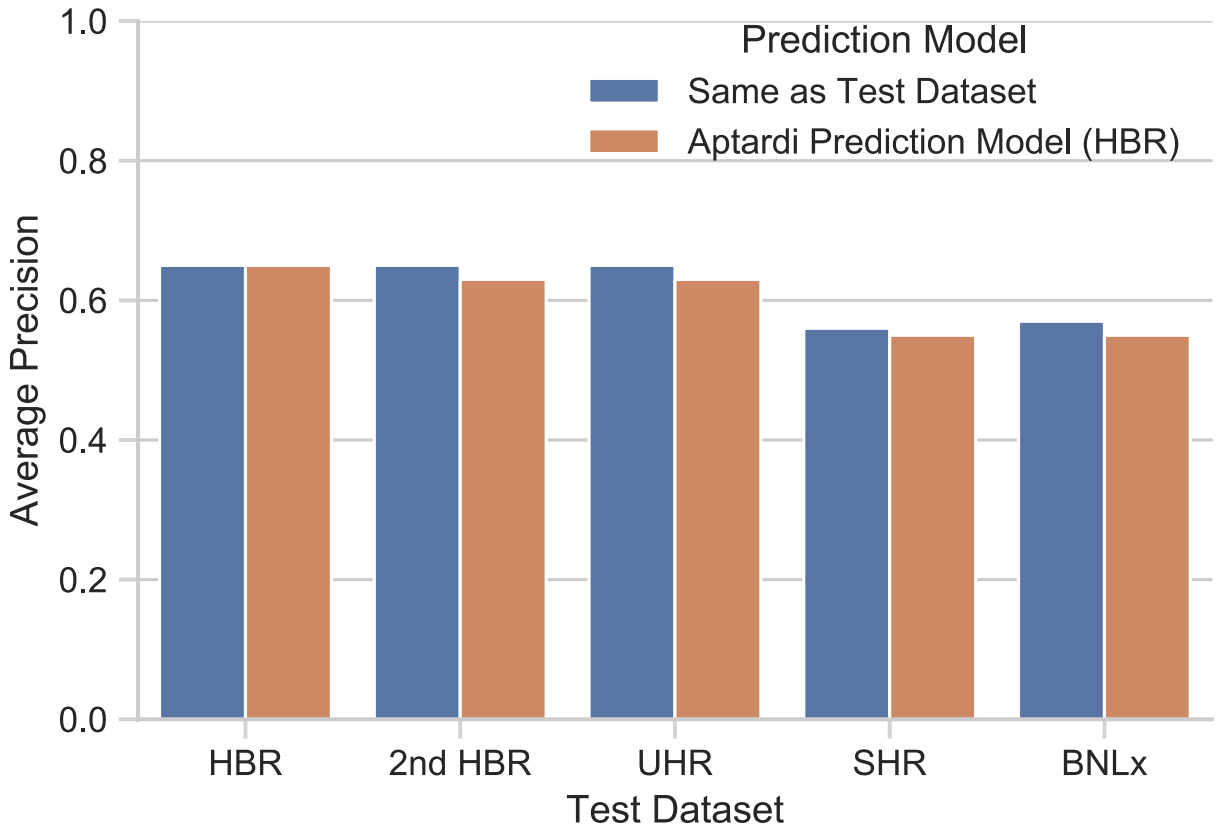



**Figure 3.3. The machine learning pipeline used to build aptardi is robust to different datasets and the aptardi prediction model generated from the Human Brain Reference dataset is applicable across diverse datasets.** Blue bars indicate the performance of the dataset-specific prediction model on its own dataset, i.e., the model was built and evaluated on a single dataset. Orange bars represent the performance of the aptardi prediction model – built from the Human Brain Reference dataset – on the given dataset (x-axis).

Also for the four alternative datasets, we built dataset-specific prediction models and compared their performance on their own dataset to the performance of the HBR-based aptardi prediction model on the given dataset to demonstrate the robustness of the machine learning pipeline used to build the aptardi prediction model. For all four datasets, the increase in AP when the same dataset for training the prediction model is used for evaluating the prediction model (as opposed to the performance of the HBR aptardi prediction model on the same given dataset) was minimal, i.e., less than or equal to 2 percentage points (Figure 3.3). Furthermore, the similarity (within two percentage points) of the AP between the training, testing, and analysis (i.e., not

merging transcripts; see Methods) sets demonstrate the aptardi prediction model is not prone to overfitting and, when these intra- and inter- model/dataset comparisons with the HBR dataset were extended to each of the four datasets, similar results were achieved (Supplementary Figure 2).

**Performance of aptardi on other true polyadenylation sites datasets**

Other polyA site datasets in the literature report polyA sites aggregated from multiple samples and sources, such as PolyASite 2.0[404] and PolyA_DB 3[405]. While aptardi was designed to make sample-specific predictions, we evaluated the performance of the aptardi prediction model (built from HBR) on these more extensive true polyA sites datasets using the HBR RNA-Seq data and hg38/GRCh38 human genome. For each polyA site cluster listed by PolyASite 2.0, the representative polyA site was used. PolyASite 2.0 and PolyA_DB 3 contained 569,005 and 289,998 annotated polyA sites leading to 102,774 and 86,685 polyA site bins, respectively (compared to 42,977 polyA site bins using the HBR PolyA-Seq data) out of the 778,166 bins produced from the transcript processing of HBR. As expected, the precision increased (PolyASite 2.0 = 0.85, PolyA_DB_3 = 0.88) and the recall decreased (PolyASite 2.0 = 0.18, PolyA_DB = 0.22) compared to the aptardi prediction model's performance on its testing dataset (precision = 0.71, recall = 0.41).

**Improvement of 3' end annotation in the transcriptome map by aptardi**

The overarching goal of aptardi is to yield an updated, sample/experiment-specific transcriptome map from the original transcriptome with more accurately annotated 3' ends of expressed polyadenylated transcripts. As such, aptardi was primarily benchmarked by comparing how it improved upon the reconstruction generated by the popular assembler StringTie[191]. The StringTie assembly also incorporated Ensembl[382] (v.99) annotation, which helps guide its

reconstruction – especially at 3’ ends. Note that aptardi outputs all original transcript structures from the input transcriptome (i.e., original transcriptome) in addition to those annotations identified by the program.

Of the 113,923 transcripts in the original transcriptome from the HBR sample (i.e., StringTie used the HBR sample RNA-Seq data and existing Ensembl annotation to generate the original transcriptome), only 39,842 (35%) had a 3’ terminus that corresponded to a true polyA site (+/- 100 bases). When the aptardi prediction model was incorporated, 27,853 transcript annotations were added to the original transcriptome where the polyA site/3’ terminus differed from its original transcript structure. Of these additional 27,853 transcripts, 22,846 (82%) matched the location of a true polyA site in the HBR PolyA-Seq data (+/- 100 bases), meaning the majority of aptardi transcript structures incorporated into the original transcriptome had accurate polyA site annotation (Figure 3.4). Furthermore, the confusion matrix of predictions made by the aptardi prediction model on each 100 base increment (i.e., bin) improved the true positive to false positive ratio compared to the original transcriptome (produced by StringTie) while simultaneously decreasing the number of false negatives in favor of true negatives (Supplementary Figure 3).

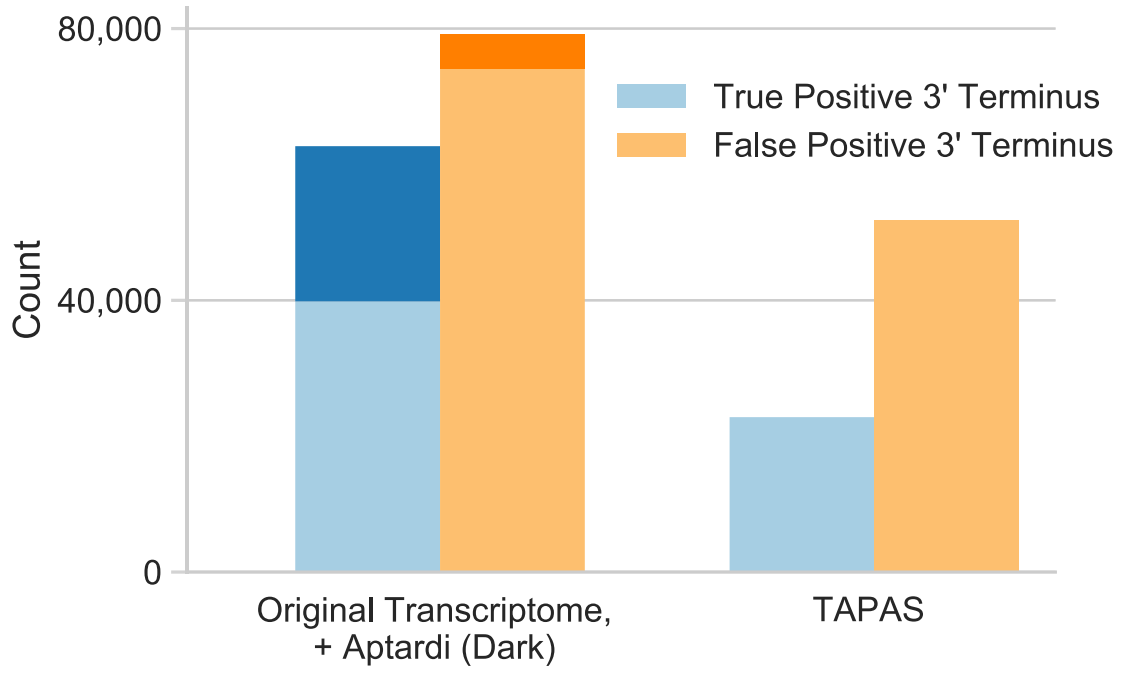



**Figure 3.4. Incorporating aptardi transcripts into the original transcriptome improves the ratio of true positive to false positive 3' termini compared to the original transcriptome and compared to the Tool for Alternative Polyadenylation site AnalysiS (TAPAS) analysis on the original transcriptome.** Results from transcripts added by aptardi to the original transcriptome are shaded in dark. Transcripts whose 3' terminus was plus or minus 100 bases of a true polyadenylation site from PolyA-Seq data were considered a true positive and otherwise counted as a false positive. Data shown are from the Human Brain Reference dataset.

We next compared aptardi to TAPAS[207] – identified by Chen et al.[197] as the top performer for characterizing APA from RNA-Seq – using the same 100 base distance cutoff to define true positives (i.e., if a TAPAS prediction was within 100 bases of any true polyA site in the HBR PolyA-Seq data it was considered a true positive, otherwise it was a false positive). The aptardi pipeline identified the 3' termini correctly for 62,688 transcripts compared to 3' termini of 22,804 transcripts using TAPAS. Although the number of transcripts with a false positive 3' terminus was higher in the aptardi modified transcriptome compared to TAPAS [207] due to annotations from the original transcriptome, the positive predictive value was higher for the aptardi modified transcriptome because it added many more true positive than false positive 3' termini to the original transcriptome (aptardi modified transcriptome = 0.44, TAPAS = 0.31).

Finally, the aptardi pipeline captured more unique true polyA sites (as identified by the HBR PolyA-Seq data) compared to both TAPAS and the original transcriptome (aptardi modified transcriptome = 29,327, TAPAS = 25,180, original transcriptome = 23,685). Similar results were achieved when adjusting the base distance cutoff for true positives and/or utilizing the PolyASite 2.0 and PolyA_DB databases to define true polyA sites (Supplementary Table 2).

A final comparison was made to APARENT [399], which utilizes only DNA sequence to make predictions. Its positive predictive value was lower than that for aptardi for all databases at all base distance cutoffs (Supplementary Table 2). Notably, while APARENT annotated more true polyA sites as defined in the PolyASite 2.0 and PolyA_DB databases compared to all the other methods (likely due to its high number of overall predictions), this increase was not observed when being compared to aptardi and using the HBR PolyA-Seq data to define the true polyA sites, which likely more accurately represents the polyA sites being expressed in the corresponding HBR RNA-Seq data (Supplementary Table 2).

**Aptardi identifies sample-specific transcripts missed by current transcriptome reconstruction methods**

We next sought to ascertain if aptardi could identify APA transcripts observed in a previous study where differential APA expression was induced by knocking down the cleavage and polyadenylation machinery CFIm25 [400]. In this study, the authors experimentally confirmed expression of short APA transcript isoforms after CFIm25 knockdown for three genes capable of undergoing APA [406, 407] – *CCND1*, *DICER1*, and *TIMP2* – and used DaPars [205] to computationally estimate the locations of polyA sites.

For each the control and knockdown RNA-Seq dataset, the aptardi modified transcriptome was compared to the original transcriptome, which contained both Ensembl

annotations and sample-specific expressed transcripts identified through StringTie reconstruction. In the control RNA-Seq dataset, neither aptardi nor the original transcriptome identified a shorter APA transcript for *CCND1* in agreement with the original study design; in the knockdown treatment RNA-Seq dataset, only aptardi recapitulated the short APA isoform (Figure 3.5a), demonstrating its sensitivity to sample-specific data and its ability to improve upon current annotation methods. Likewise, only aptardi identified the proximal APA transcript for *DICER1* (Figure 3.5b). For *TIMP2*, multiple transcript isoforms are annotated in Ensembl[382], and StringTie[191] retained all these transcripts in its reconstruction. In contrast, aptardi annotated a short APA transcript only in the treatment consistent with Masamha et al.[400], again demonstrating its sample-specific sensitivity (Figure 3.5c). Finally, the locations of the proximal transcripts for these genes identified by aptardi were similar to the original study (*CCND1*: aptardi = chr11: 69,651,917, original study = chr11: 69,651,578; *DICER1*: aptardi = chr11: 95,090,264, original study = chr11: 95,090,400; *TIMP2*: aptardi = chr17: 78,855,465, original study = chr17: 78,855,601).

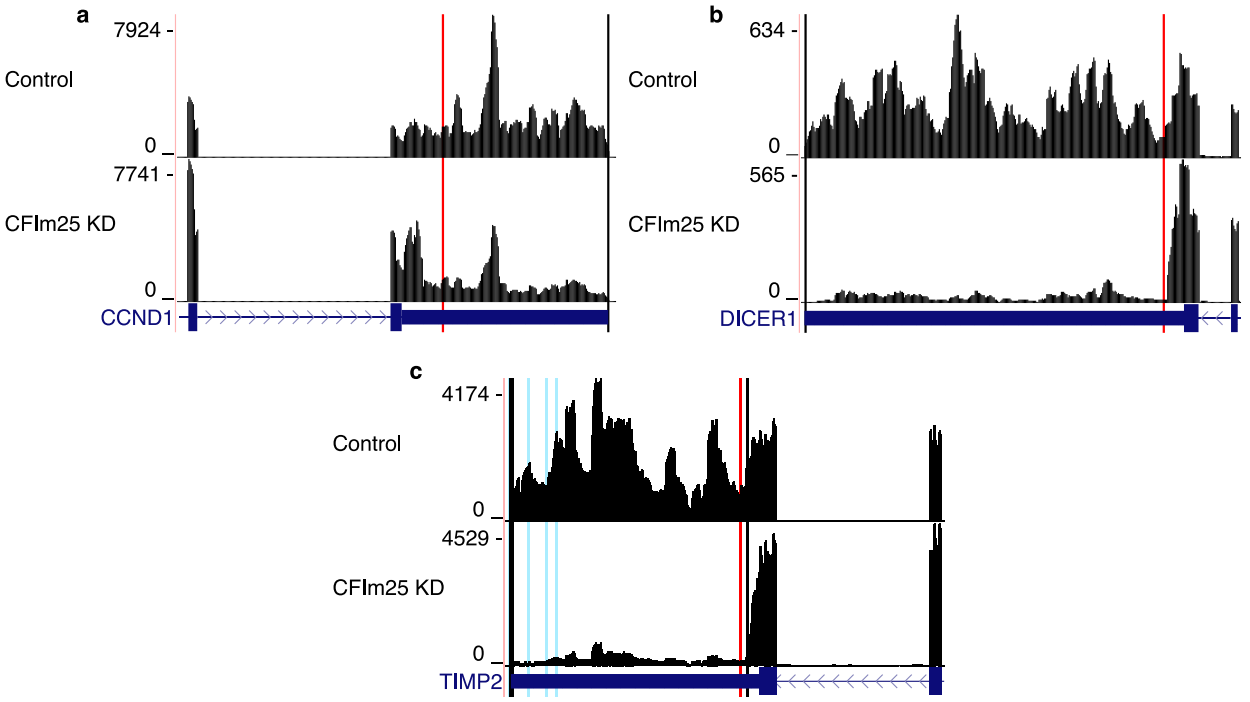



**Figure 3.5. Aptardi displays sample specific sensitivity when annotating transcription stop sites.** RNA sequencing read densities for **a,** *CCND1*, **b,** *DICER1*, and **c,** *TIMP2* after control (Control) siRNA treatment and CFIm25 knockdown (KD) in HeLa cells. Numbers on y-axis indicate RNA-Seq read coverage. After knockdown, each gene preferentially expresses a proximal alternative polyadenylation (APA) site compared to under control conditions. Transcript structures shown are from RefSeq annotation (dark blue), where boxes and lines indicate exons and introns, respectively. Black vertical lines indicate transcript stop sites identified in the original transcriptome, red vertical lines indicate transcript stop sites only identified in the aptardi modified transcriptome and that match the original study's findings, and blue vertical lines indicate transcript stop sites only identified in the aptardi modified transcriptome that are not described in the original study. Graphics were generating using the UCSC Genome Browser (https://genome.ucsc.edu/) using the hg38 human genome assembly.

**Comparison of aptardi predictions across mouse tissues**

The above study demonstrated aptardi's ability to differentiate polyA sites across samples; however, we extended this analysis by comparing different mouse tissues – namely liver and brain – to mimic subtle differences in polyA sites. Brain and liver RNA-Seq data were procured from Li et al. [401], true polyA sites were derived again from PolyA-Seq data that were specific to brain and liver, and the mouse reference mm10/GRCm38 genome was used for DNA sequence for both tissues (see Mouse tissue analysis in Methods for more details). For the

PolyA-Seq data, we identified polyA sites that 1) were unique to either brain or liver, 2) were within the 3' modified terminal exon of a transcript in the StringTie generated original transcriptome (i.e., made available to aptardi) for both tissues, and 3) did not coincide with a previously annotated Ensembl transcript end. Using these restrictions, we were able to focus on unannotated polyA sites that differed across tissues but were associated with genes/transcripts expressed in both tissues. This resulted in 756 unannotated brain-specific sites and 1,529 unannotated liver-specific sites. The StringTie pipeline was able to capture three of the unannotated brain-specific sites and four of the unannotated liver-specific sites. Including the aptardi prediction model (built from HBR) in the transcript discovery pipeline added 26 unannotated brain-specific sites (fold increase = 9) and 69 unannotated liver-specific sites (fold increase = 17) Furthermore, only nine of these 26 additional unannotated brain-specific polyA sites were also added to the liver data using aptardi (i.e., aptardi did not distinguish the brain from liver) and only 25 of these 69 unannotated liver-specific polyA sites were identified by aptardi in the brain data.

**Evaluation of the use of aptardi in a differential expression pipeline**

The influence of aptardi on differential expression analysis was evaluated using the BNLx and SHR rat brain datasets by evaluating transcripts identified as differentially expressed between strains with the aptardi modified transcriptome (p-value <= 0.001) but not the original transcriptome derived from the StringTie/Ensembl pipeline (p-value >= 0.001). Note that since aptardi incorporates transcripts into annotation, expression levels of existing transcripts can also change, i.e., transcripts present in both the aptardi modified transcriptome and the original transcriptome may be identified as differentially expressed in one and not the other. A total of 1,166 out of 32,348 transcripts and 918 out of 28,329 transcripts expressed above background

were differentially expressed (p-value <= 0.001) using the aptardi modified transcriptome and original transcriptome, respectively. A total of 40 transcripts that could be associated with an Ensembl gene symbol were differentially expressed in the aptardi modified transcriptome but not in the original transcriptome although they had identical structures, including 3' ends, in both (i.e., original transcriptome transcripts). Furthermore, 54 aptardi transcripts that could be associated with a gene symbol were differentially expressed and NOT measured/identified in the original transcriptome (Supplementary Data 1). The RNA-Seq read coverage for six of these aptardi transcripts are depicted in Figure 3.6. For *Unc79* (Figure 10a), *Sf3b1* (Figure 3.6b), *Ptn* (Figure 3.6c) and *Ap3b1* (Figure 3.6d) the original transcript was differentially expressed in the aptardi modified transcriptome but not the original transcriptome, and for *Zdhhc22* (Figure 3.6e) and *RGD1559441* (Figure 3.6f) the aptardi transcript was differentially expressed. The RNA-seq read coverage across these genes support the presence of the aptardi transcripts and differential expression of the various isoforms between strains. Moreover, these results demonstrate that aptardi is capable of identifying both shortening and lengthening events – e.g., four of the six genes were annotated with a shorter transcript by aptardi and two of the six a longer one – as well as identifying isoforms across a broad range of RNA-Seq coverage depths; the peak coverage value for each gene ranged from approximately 200 to 8,000.

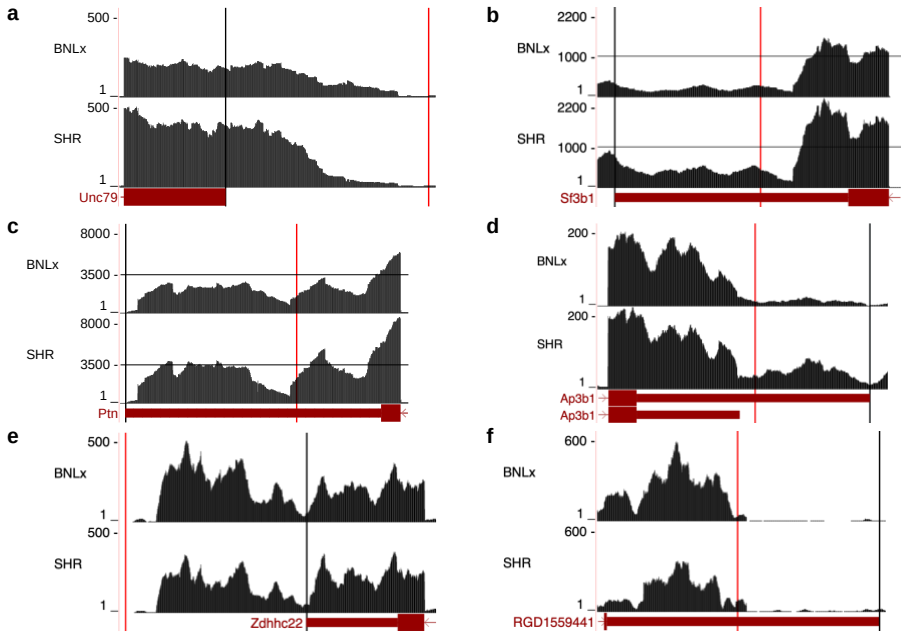



**Figure 3.6. Incorporation of aptardi into differential expression analyses.** RNA sequencing read densities for six genes in BNLx and SHR inbred rat strains. Numbers on y-axis indicate RNA-Seq read coverage. Read coverage represents the aggregate of three biological samples for each strain. Transcript structures shown are from Ensembl annotation (dark red), where boxes and lines indicate exons and introns, respectively. Black vertical lines denote transcript stop sites identified in the original transcriptome derived using StringTie, and red vertical lines indicate transcript stop sites identified in the aptardi modified transcriptome only. No transcripts were identified as differentially expressed between strains in the original transcriptome ($p > 0.001$), but at least one differentially expressed transcript for each gene was identified in the aptardi modified transcriptome ($p <= 0.001$). For **a,** *Unc79* **b,** *Sf3b1* **c,** *Ptn* and **d** *Ap3b1* the original transcript isoform (black line) was differentially expressed in the aptardi modified transcriptome, and for **e** *Zdhhc22* and **f** *RGD1559441* the aptardi transcript was differentially expressed (red line). Graphics were generating using the UCSC Genome Browser (https://genome.ucsc.edu/) using the rn6 rat genome assembly.

## Discussion

Aptardi leverages the information afforded by both DNA sequence and short-read RNA-Seq to accurately annotate the polyA sites of expressed transcripts in a biological sample. We first established the applicability of aptardi by showing that 1) a prediction model derived from a single dataset performed well on datasets that differ on technical issues and even species, 2) the process of training the prediction model is generalizable across different types of RNA-Seq data/DNA sequence, and 3) the algorithm is not prone to overfitting. Namely, we showed that the

aptardi prediction model provided for users (built from the HBR dataset) performs equally well on RNA-Seq datasets derived from different library preparations, organisms, and with different RNA sequencing depths. We note that aptardi performed modestly worse on the BNLx and SHR datasets and hypothesize this is because the true polyA sites were derived from the Sprague Dawley rat instead of the specific rat strain. This is supported by the fact that prediction models built from the BNLx and SHR datasets and tested on these same datasets performed similarly to the aptardi prediction model built on the HBR dataset (Supplementary Figure 2). However, we cannot rule out the possibility that this is due to the unstranded RNA-Seq for these datasets and/or different species. Moreover, the comparable results when models were built and evaluated on a single dataset versus models applied to different datasets from those used to train the model suggest the data processing pipeline/prediction models are generalizable. Finally, the similarity of the average precision estimates between training and testing sets demonstrates that aptardi is not prone to overfitting. Overall, these results indicate aptardi can be broadly used.

We next established that incorporating aptardi into current transcriptome reconstruction methods improves annotation of 3' ends. This was done by comparing the aptardi modified transcriptome to the original transcriptome assembled using the power of both existing annotation via Ensembl and taking into consideration RNA-Seq coverage via StringTie. Adding aptardi transcripts increased the number of unique true polyA sites captured by the transcriptome and furthermore increased the ratio of true positive to false positive termini compared to the original transcriptome. Aptardi also outperformed TAPAS in these respects, and TAPAS was previously identified as the top performer for identifying polyA sites from RNA-Seq [197]. Likewise, aptardi produced greater positive predictive values compared to the deep learning algorithm APARENT that utilizes DNA sequence to make predictions and, furthermore,

annotated more polyA sites at 100 and 50 base distance cutoffs and nearly the same at a base distance cutoff of 25 despite making many fewer predictions.

Applying aptardi in control and CFIm25 knockdown RNA-Seq data demonstrated its 1) sensitivity to sample-specific expression, 2) ability to identify both shortening and lengthening APA events, 3) competence across a broad range of RNA-Seq coverage depths, and 4) ability to improve upon current reconstruction methods. Of interest, in the control for *TIMP2*, aptardi identified several 3' ends close to the annotated distal transcript that were not noted in the original study[400]. The RNA-Seq from the control sample displays uneven coverage in this region, meaning aptardi may have uncovered additional, previously unknown isoforms (Figure 3.5c). Of note, the library preparation for these RNA-Seq data were unstranded (unlike the data used to generate aptardi), further supporting aptardi's broad applicability. These results also highlight potential weaknesses of aptardi. For instance, aptardi incorporates a single 3' end into multiple transcripts for each gene listed because it does not distinguish transcripts overlapping the same genomic region. This may be somewhat mitigated by curating a more selective input transcriptome. Here the entire Ensembl annotation was provided, which includes many "pseudo" transcripts (e.g., retained introns, nonsense mediated decay, and isoforms only identified computationally), and these transcripts often overlap manually identified mRNAs. Secondly, aptardi will add polyA sites for a transcript regardless of if another transcript isoform from the same gene already has a transcript stop site at the given location, as is the case for *TIMP2*. As a result, it is possible that aptardi incorporates a transcript stop site belonging to a different transcript.

The sample-specific sensitivity of aptardi was also evaluated using RNA-Seq from brain and liver mouse tissues and evaluating its ability to identify polyA sites unique to each tissue.

Aptardi successfully annotated tissue-specific polyA sites not previously reported in Ensembl annotation or discovered by StringTie reconstruction. Furthermore, aptardi was able to provide a significant fold increase in polyA site annotation compared to StringTie, exemplifying its sensitivity. While the overall number of polyA sites detected was modest, this was likely due to the low sequencing depths of the samples (Supplementary Table 3). As a result, many polyA sites reported in the PolyA-Seq data – which was sequenced at much greater depths – were likely not detectable here. Indeed, the reference mouse Ensembl genome annotated a greater number of both overall polyA sites (brain = 14,630, liver = 13,209) and tissue-specific polyA sites (brain = 5,144, liver = 3,668) compared to the StringTie pipeline (overall: brain = 11,285, liver = 9,572; tissue-specific: brain = 3,468, liver = 2,331). StringTie removes guide transcripts (in this case Ensembl transcripts) below a minimum coverage threshold (one when using default settings), suggesting many PolyA-Seq polyA sites – including those in the 3' modified terminal exons that aptardi did not annotate – were not detectable here.

We further examined how incorporating aptardi into downstream transcriptome analyses such as differential expression may alter interpretation of results. We found that multiple isoforms – some of which were already present in the original transcriptome and some of which were transcripts identified by aptardi – were differentially expressed between BNLx and SHR recombinant inbred rats only when using the aptardi modified transcriptome. Some of these transcripts are derived from genes that have also been implicated in phenotypes related to the SHR rat, such as greater sensitivity to addictive drugs [408] and increased voluntary ethanol consumption [409]. For instance, expression of *Ptn* – which has verified APA sites [410] – is modulated by amphetamine in rat nucleus accumbens [411]. Furthermore, *Unc79* knockout mice displayed hypersensitivity to ethanol, e.g. increased preference for and consumption of alcohol

[409]. Undoubtedly further investigation is needed to elucidate the role of *Ptn and Unc79* APA on addiction phenotypes, but these preliminary results demonstrate how aptardi may help unravel the genetic architecture of complex diseases. Of note, while Ensembl annotation provides two transcript isoforms for *Ap3b1*, StringTie assembly resulted in inclusion of the longer isoform only, highlighting the difficulty of current assembly methods for identifying 3' ends of transcripts embedded in a longer version. However, aptardi identified a shorter transcript within 100 bases of the original Ensembl annotation for the shorter transcript that is likewise supported by the RNA-Seq data (Figure 3.6d).

Also of note, aptardi is easily integrable into existing analyses pipelines. For instance, unlike current supplemental methods designed for APA detection from RNA-Seq, no additional data manipulation is required prior to running the program. The input files are readily available (e.g. reference genome and reference transcriptome) or already generated during the course of transcriptomic analysis (e.g. RNA-Seq data and a reconstructed transcriptome). The output GTF file can be used in the same manner as other annotation files (e.g. those accessed via Ensembl or generated via a transcriptome assembler such as StringTie). Moreover, the program can be seamlessly integrated into a single operation with upstream transcriptome assembly and downstream analyses (e.g. quantitation) via piping to make for streamlined analysis. Finally, there is also the option of constructing a prediction model using the aptardi architecture, which increases the breadth of its applicability to diverse data sources.

Transcriptome profiling is one of the most utilized approaches for investigating human diseases at the molecular level, yielding important insights into many pathologies. A prerequisite for these studies is a representative transcriptome map. Aptardi incorporates APA transcripts to produce a more accurate transcriptome map, thereby enabling future research into the role of

APA transcripts – as well as other transcripts unencumbered by convoluted annotation with APA transcripts – in human health and disease.

Aptardi is implemented in Python and is freely available as open source software (https://github.com/luskry/aptardi).

# CHAPTER IV

# BEYOND GENES: INCLUSION OF ALTERNATIVE SPLICING AND ALTERNATIVE POLYADENYLATION TO ASSESS THE GENETIC ARCHITECTURE OF PREDISPOSITION TO VOLUNTARY ALCOHOL CONSUMPTION IN BRAIN OF THE HXB/BXH RECOMBINANT INBRED RAT PANEL

## Introduction

Post transcriptional phenomena are powerful mechanisms by which eukaryotes expand their genetic diversity. For instance, researchers estimate that 95-100% of multi-exon genes in human can undergo alternative splicing [412-414]. Likewise, an estimated 70% or more mammalian genes have multiple polyadenylation sites [208, 415] and can therefore express alternative polyadenylation isoforms.

Biologically, alternative splicing can lead to diverse functions [416] by changing the protein encoded by the mRNA [417]. In contrast, the vast majority of alternative polyadenylation occurs in the 3' UTR [102] and thus generates identical proteins. Regardless, alternative polyadenylation profoundly impacts the mRNA by modifying its stability, translocation, nuclear export, and cellular localization, as well as the localization of the encoded protein [102, 103]. Alternative polyadenylation most often exerts its effects though gain or loss of miRNA binding sites the 3' UTR; more than 50% of conserved miRNA binding sites reside downstream of the most proximal polyadenylation site in mammalian genes [418].

Alternative splicing and alternative polyadenylation have increasingly been associated with disease. For example, alternative splicing has been recognized as a genetic modifier of disease phenotype [416] and susceptibility to disease [419]. Notably, alternative splicing has been shown to impact the phenotypic variation of diseases, or in other words impact quantitative

(complex) traits, via changes in the expression of alternative "normal" transcripts or in the relative pattern of different mRNA isoforms [420]. One example of the latter is the tau protein; exon 10 can be included or skipped, and dysregulation of the ratio between these two alternative splicing isoforms can lead to the development of inherited frontotemporal dementia and parkinsonism linked to chromosome 17 [421-423]. Several other alternative splicing events have been linked to neurological diseases [424].

Although alternative polyadenylation is a relatively new research area, it too has been associated with biological processes such as the innate antiviral immune response, cancer initiation and prognosis, and developing drug resistance [418]. Similarly, differences in expression of alternative polyadenylation transcripts has been implicated in disease [169] and alternative polyadenylation is increasingly being acknowledged as a risk factor for complex diseases [170].

The objective in our present study was twofold: 1) characterize the alternative splicing and alternative polyadenylation transcriptional landscape in brain of the HXB/BXH RI rat panel, and 2) identify candidate transcripts associated with the complex trait of voluntary alcohol (i.e., ethanol) consumption using both a network and individual transcript approach. Throughout this paper, we use the term 'transcript' to refer to the sequence of a processed RNA and the term 'gene' as the DNA locus that the transcript is transcribed from. In this context a gene can produce many distinct transcripts through alternative splicing and alternative transcription start and stop sites (e.g., alternative polyadenylation).

To accomplish our first goal, we procured an extensive expression library of whole brain in the HXB/BXH RI panel consisting of over one terabyte of RNA sequencing (RNA-Seq) data. We then applied two computational methods, StringTie [191] and aptardi [425], to characterize

*in vivo* expression of alternative splicing and alternative polyadenylation in the brains of these animals. We furthermore developed a transcriptome generation pipeline that integrates these tools and also filters transcripts to eliminate potential false positives yielding a high-quality transcriptome.

To accomplish our second goal, we integrated expression data with phenotypic data to identify candidate coexpression networks and individual candidate transcripts that predispose these rats to voluntary alcohol consumption, i.e., expression data were taken without exposure to alcohol, similar to our previous work [129, 132, 138, 240]. Particularly, we previously used whole brain expression data and this phenotype to identify candidate gene coexpression networks [132]. However, that study utilized expression data in the form of microarrays and analysis was limited to genes and transcripts that were unambiguously probed by the microarray. In the current study, we sought to go beyond genes and identify specific transcripts associated with the trait by harnessing the power of our deep RNA-Seq libraries in these animals.

## Materials and Methods

### Animals

The HXB/BXH RI rat panel, a subset of the Hybrid Rat Diversity Panel, was used in this study. This RI panel consists of 30 strains derived from the congenic Brown Norway strain with polydactyly-luxate syndrome (BN-Lx/Cub) and the Wistar origin spontaneously hypertensive rat strain (SHR/OlaIpcv) using gender reciprocal crossing and more than 80 generations of brother sister mating after the F2 generation [248].

### Voluntary Alcohol Consumption in the HXB/BXH Recombinant Inbred Rat Panel

Voluntary alcohol consumption was measured using a two-bottle choice paradigm and 23 HXB/BXH strains and the two progenitor strains of the RI panel. Specifically, rats were provided

10% ethanol as their only fluid during week zero. During weeks 1-7, rats were given a two-bottle choice between water and ethanol. The average daily alcohol consumption (g/kg body weight) during week two represents the voluntary alcohol consumption phenotype for this analysis (Figure 4.1).

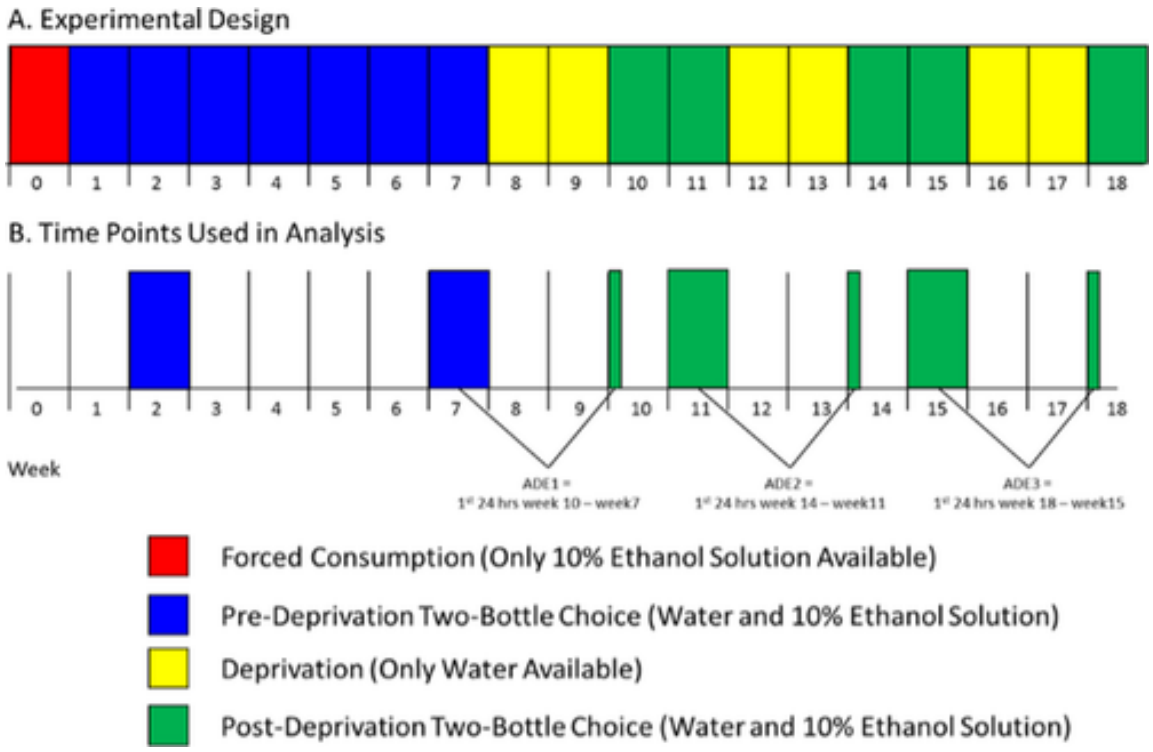



**Figure 4.1. From [129].** The experimental design for the voluntary alcohol consumption phenotype used in this study. Measurements from week two were used.

This phenotype was described in our previous study [129], has been used in previous genetics studies [129, 132, 426], and has an established its heritability ($R^2 = 0.39$) [129, 426], making it amenable to this genomics study. Only voluntary alcohol consumption data from HXB/BXH strains with corresponding RNA expression (21 strains) and/or genetic marker data (21 strains) were used (Supplementary Table 1). Moreover, these data are publicly available through PhenoGen [132] (http://phenogen.org).

**Whole Brain RNA Sequencing**

The University of Colorado Anschutz Medical Campus received shipments of brain tissue from male rats (~70-90 days old) stored in liquid nitrogen from Dr. Michal Pravenec at the Institute of Physiology of the Czech Academy of Sciences. These studies were performed in accordance with the Animal Protection Law of the Czech Republic and were approved by the Ethics Committee of the Institute of Physiology, Czech Academy of Sciences, Prague. A total 90 HXB/BXH biological replicates (i.e., brains from individual rats) from 30 strains were received, as well as 3 SHR/OlaIpcv progenitor strain samples (93 samples total). The SHR/OlaIpcv samples were used as a loading controls and eight, two, or one technical replicate(s) were generated from each biological sample. Including the technical replicates for the SHR strains, 101 RNA-Seq libraries were generated.

Total RNA (>200 nucleotides) was extracted from whole brain using the RNeasy Plus Universal Midi Kit (Qiagen, Valencia, CA, USA) and cleaned using the RNeasy Mini Kit (Qiagen, Valencia, CA, USA). Four μL 1:100 dilution of either ERCC Spike-In Mix 1 or Mix 2 (ThermoFisher Scientific, Wilmington, DE, USA) was added to each RNA sample. The Illumina TruSeq Stranded RNA Sample Preparation kit (Illumina, San Diego, CA, USA) was used to construct sequencing libraries, which included ribosomal RNA depletion using the Ribo-Zero rRNA reduction chemistry. Sequencing library quality was evaluated using an Agilent Technologies Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA). Samples were sequenced in eight batches on an Illumina HiSeq2500 or HiSeq4500 (Illumina, San Diego, CA, USA) in High Output mode to generate 2X100 or 2X150 paired end reads.

**Brain Specific Transcriptome Generation and Quantitation Using Whole Brain RNA Sequencing**

Reference annotation, especially in non-model organisms such as rat, lack annotation of alternative splicing and alternative polyadenylation transcripts. Likewise, reference annotation represents data pooled from multiple tissues, experiments, etc., and thus may not accurately represent the transcriptome for a given study. Therefore, we sought to generate a brain specific transcriptome map of the HXB/BXH RI panel used in this study by incorporating RNA expression data (in the form of short read RNA-Seq), DNA sequence information, and computational methods, namely StringTie and aptardi, to annotate expressed alternative splicing and alternative polyadenylation transcripts, respectively. To identify high confidence transcripts, we furthermore quantitated transcripts and removed those that were lower expression. We likewise quantitated the transcriptome to enable downstream quantitative analyses evaluating the role of transcripts/genes in predisposition to voluntary alcohol consumption. An outline of the transcript generation and quantitation steps are presented in Supplementary Figure 1.

*Read Processing for Quality*

Initially, adapter sequences and low quality base calls were eliminated from raw reads from the 90 HXB/BXH RNA-Seq samples (one rat per library) and 11 SHR/OlaIpcv RNA-Seq libraries (i.e., loading controls) using cutadapt (v.1.9.1) [377]. Reads aligning to rRNA from the RepeatMasker database [427] (accessed through the UCSC Genome Browser; https://genome.ucsc.edu/) [381] using Bowtie 2 (v.2.3.4.3) [428] were removed.

*Evaluation of Unannotated Genes and Unannotated Splicing – StringTie Transcriptome Generation*

RNA-Seq libraries from each sample were aligned to their strain-specific genomes using HISAT2 (v.2.1.0) [429], and then alignments from the same strain were concatenated across biological replicates using SAMtools (v.1.9) [379] *merge* to generate a single genome alignment per strain. Strain-specific genomes were constructed from the Rat Genome Sequencing Consortium (RGSC) Rnor_6.0/rn6 version of the rat genome [430] by imputing single nucleotide polymorphism (SNP) information for each strain based on their STAR Consortium genotypes [431] and DNA sequencing (DNA-Seq) data from male rats of the progenitor strains. The DNA sequencing data are publicly available on the PhenoGen website [132] (http://phenogen.org). StringTie (v.1.3.5) [191] was used to generate *de novo* strain-specific transcriptomes using default settings, which used the strain-specific genome alignment and the rat Ensembl reference transcriptome (v.99) [432] to guide transcriptome assembly. A combined StringTie transcriptome for the HXB/BXH RI panel was generated using the *merge* functionality of StringTie and providing each strain-specific transcriptome and the rat Ensembl reference transcriptome as a guide.

*Evaluation of 3' Termini – Aptardi Transcriptome Generation*

The transcriptome was further processed by aptardi to identify alternative polyadenylation transcripts [425]. For aptardi (v.1.0.0) analysis, a single file with alignment of all RNA-Seq reads to the genome, the reconstructed transcriptome file, and a genome sequence file are required. To generate a single BAM file, the strain-specific BAM files of all HXB/BXH RI strains (but not the SHR/OlaIpcv BAM file) were merged using SAMtools *merge*. The combined StringTie transcriptome was used as the input GTF reconstructed transcriptome file,

and the rat Rnor_6.0/rn6 reference genome accessed via the UCSC Genome Browser [430, 433] was used for genomic sequence information.

*Detected Above Background Transcriptome Generation and Quantitation*

After analyzing the transcriptome for alternative splicing and alternative polyadenylation transcripts of genes, quantitation was used to establish the detected above background (DABG) transcriptome. Prior to quantitation, transcripts not derived from autosomal or sex chromosomes were removed. Transcripts in the aptardi transcriptome were then quantitated in each of the 90 HXB/BXH RNA-Seq samples using RSEM (v.1.3.0) [257]. Transcripts with zero estimated read counts in one third or more samples were removed, as well as transcripts 200 nucleotides or fewer in length. This high-quality transcriptome was then used to re-quantitate transcripts with RSEM for each of the HXB/BXH samples as well as the 11 SHR/OlaIpcv samples. Prior to transcript filtering, libraries with less than 10 million paired end reads were removed from all subsequent analyses, resulting in the removal of a single SHR/OlaIpcv sample. Transcripts with zero counts in one third or more of samples were again removed to yield the DABG transcriptome and the corresponding estimated read counts for each RNA-Seq library and for each transcript.

*Quantitation Normalization for Weighted Gene Coexpression Network Analysis and Correlation Analysis*

Estimated read counts of transcripts in the DABG transcriptome for each RNA-Seq library generated by RSEM were normalized for sequencing depth using upper quartile normalization [434] implemented in EDASeq (v.2.22.0) [435] followed by a regularized log (rlog) normalization with DESeq2 (v.1.28.1) [436].Finally, expression values were adjusted for batch effects using ComBat [437] from sva (v.3.36.0) [438].

## Quantitative Trait Loci Analysis

### *Genetic Markers for Quantitative Trait Loci Analyses*

Genetic makers were initially procured from publicly available SNP genotype data for these rats originally obtained by the STAR Consortium [431]. Probes from the original array were aligned to the Rnor_6.0/rn6 version of the rat genome using BLAT [439]. Markers were further processed into unique strain distribution patterns for QTL analyses as detailed in our previous work [440]. These data are publicly available on the PhenoGen website (http://phenogen.org), including the version specific to this paper.

### *Statistical Methods for Quantitative Trait Loci Analyses*

Quantitative trait loci (QTL) analysis was performed for the behavioral phenotype (pQTL), for module eigengenes (meQTL), and for transcript expression levels (eQTL). Marker regression was using for all QTL analyses. Likewise, all empirical genome-wide p-values were calculated using 1,000 permutations [441]. Both significant (genome-wide p-value < 0.05 and suggestive (genome-wide p-value < 0.63) QTL were considered for pQTL analysis [442, 443]. For meQTL and eQTL analyses, a stricter genome-wide significant p-value of < 0.01 was enforced. Strain mean voluntary alcohol consumption values were used for pQTL analysis, strain mean transcript normalized expression estimates were used for eQTL analyses, and module eigengene values (which were produced from WGCNA using strain mean transcript normalized expression estimates and thus represent strain means) were used for meQTL analysis. The 95% Bayesian credible intervals of significant or suggestive pQTL, significant meQTL, and significant eQTL were estimated and all QTL analyses and graphics were generated using the R/qtl package (v.1.47-9) [444].

## Heritability of Transcripts

Heritability of transcripts in the DABG transcriptome was estimated as the R-squared value from a one-way ANOVA of individual rat expression values using strain as the predictor and transcript normalized expression estimates as the response.

## Identification of Candidate Coexpression Networks and Candidate Individual Transcripts Associated with Voluntary Alcohol Consumption

To be included in association analyses with alcohol consumption, transcripts needed to be one of the three expressed isoforms of a gene with the highest expression levels where expression levels were represented by the mean transcripts per million (TPM) value (as determined vis RSEM) across individual rat RNA-Seq samples for rats with voluntary alcohol consumption data only (63 samples, 21 strains, see Supplementary Table 1). Furthermore, among these, only transcripts with high heritability (greater than the median value of all transcripts in the DABG transcriptome) were included to focus on genetically influenced transcripts. Finally, transcripts that could be associated with a gene symbol through Ensembl were included for interpretability. A gene name is assigned to an individual transcript if any of the other transcripts associated with that same gene share a splice junction with an annotated gene. In this context, the assumption is that unannotated transcripts are novel splice variants of an annotated gene. For the candidate individual transcripts, there was no requirement that the eQTL must be local – that is, the location of the eQTL was not required to reside near the location of the transcript itself.

### *Weighted Gene Coexpression Network Analysis*

The WGCNA R package (v.1.69) was used to build a transcript coexpression network and to identify coexpression modules within that network from the strain mean normalized

expression estimates of transcripts. Minimum module size was set five and the deepSplit parameter was set to four to promote identification of smaller modules, but otherwise default settings were used. The soft-thresholding index (β) was set to seven to approximate scale-free topology [158] in an unsigned network (Supplementary Figure 2). The module eigengene (first principal component) was used to summarize transcript expression values within a module across strains [160].

*Candidate Coexpression Networks and Individual Candidate Transcripts*

Multiple criteria similar to those previously established [440] were used to determine candidate modules and individual candidate transcripts. 1) The Spearman's rank correlation to voluntary alcohol consumption must be significant (p-value < 0.01). For the module analysis the module eigengene expression values were used for correlation analysis, and for the individual transcripts the strain mean normalized expression estimates were used. 2) The module eigengene QTL (meQTL; for modules) or expression QTL (eQTL; for individual transcripts) must have genome-wide significance (p-value < 0.01) and must overlap a significant (p-value < 0.05) or suggestive (p-value < 0.63) phenotypic QTL (pQTL) using 95% Bayesian credible intervals.

## Results

### Whole Brain RNA Sequencing

After processing reads for quality, the number of paired end reads in the 90 HXB/BXH RI panel RNA-Seq libraries ranged from 29 million to 199 million (median number of paired end reads per sample = 71 million). The number of paired end reads in each RNA-Seq library, including the 10 SHR/OlaIpcv libraries, is provided in Supplementary Table 2.

The median strain specific genome alignment rate of the 90 HXB/BXH RI panel RNA-Seq libraries was 97% (interquartile range = 96.8% to 97.5%) and the alignment rate ranged from

79% to 98%. The alignment rate of each RNA-Seq library, including the 10 SHR/OlaIpcv libraries, is provided in Supplementary Table 3.

## Brain Specific Transcriptome Generation for the HXB/BXH Recombinant Inbred Rat Panel

### *Comparison of the Number of Isoforms Per Gene in the Reference, StringTie, Aptardi, and Detected Above Background Transcriptomes*

The rat Ensembl reference transcriptome (v.99) contains 32,586 genes and 40,772 transcripts (only including genes/transcripts derived from autosomal and sex chromosomes) for a transcript:gene ratio of 1.25 (Table 4.1). Using the reference transcriptome to guide its assembly, the StringTie transcriptome yielded 33,649 genes and 83,920 transcripts. Of the 83,920 transcripts, 40,754 were from reference annotation and 43,166 were identified by StringTie (i.e., transcripts without existing annotation in Ensembl). Applying aptardi to the StringTie transcriptome resulted in an aptardi transcriptome with an additional 71,757 transcripts identified by the program (only including genes/transcripts derived from autosomal and sex chromosomes) for a total of 155,677 transcripts. After filtering to produce the final, DABG transcriptome, there were 19,517 genes and 59,751 transcripts, or a transcript:gene ratio of 3.06. Of the 59,751 transcripts, 17,028 were derived from the reference transcriptome, 24,219 were derived from StringTie, and 18,504 were derived from aptardi. For comparison, 83% of genes in the reference Ensembl transcriptome possessed a single transcript, i.e., genes without documented alternative splicing or alternative polyadenylation isoforms (Supplementary Figure 3A). In contrast, only 44% of genes in the DABG transcriptome expressed a single transcript (Supplementary Figure 3B).

**Table 4.1. Summary of the genes and transcripts at each step in the transcriptome generation pipeline.**

| Dataset | Transcriptome Generation Step | # Genes | # Transcripts | Transcript:Gene Ratio | # Reference Transcripts | # StringTie Transcripts | # Aptardi Transcripts |
|---|---|---|---|---|---|---|---|
| Reference | Step 1 | 32,586 | 40,772 | 1.25 | 40,772 (100%) | | |
| StringTie | Step 2 | 33,649 | 83,920 | 2.49 | 40,754 (48.6%) | 43,166 (51.4 %) | |
| Aptardi | Step 3 | 33,649 | 155,677 | 4.63 | 40,754 (26.2%) | 43,166 (27.7%) | 71,757 (46.1%) |
| DABG | Step 4 | 19,517 | 59,751 | 3.06 | 17,028 (28.5%) | 24,219 (40.5%) | 18,504 (31.0%) |

The percent of total transcripts identified by each source at each step is shown in parenthesis. Reference annotation represents the rat Ensembl reference transcriptome (v.99) and was used as input for StringTie, along with whole brain RNA sequencing data, to characterize alternative splicing in brain of the HXB/BXH recombinant inbred rat panel. RNA sequencing data was likewise used, in conjunction with DNA sequence of the reference Rnor_6.0/rn6 rat, to identify brain specific alternative polyadenylation events in the HXB/BXH recombinant inbred rat panel. Finally, transcripts were filtered based on their expression estimates as determined by RSEM to retain only transcripts with substantial expression in the detected above background (DABG) transcriptome.

*Evaluation of 3' Termini*

Since this is one of the first demonstrations of the integration of aptardi into the transcriptome reconstruction pipeline in a dataset of this magnitude, it was important to examine, in detail, the 3' termini of the resulting transcriptome. Aptardi was designed to identify expressed polyadenylation sites based on RNA-Seq read coverage and on signals with the DNA sequence. To integrate this information with the StringTie transcriptome, aptardi-identified polyadenylation sites were assigned to all possible transcripts with similar 3' terminal exons since aptardi uses these locations to identify polyadenylation sites but cannot distinguish the site between these

transcripts. The goal of the two-step iterative process of estimating read counts was to use the strategies implemented in RSEM to assign reads to different transcripts/polyadenylation site pairs to clarify which transcript(s) that polyadenylation site should be associated with, i.e., more reads equate to higher confidence in the transcript/polyadenylation site pair. To examine the performance of this pipeline we 1) compared the number of transcript/polyadenylation site pairs in the aptardi transcriptome to the number of pairs that remained in DABG transcriptome, i.e., could we use read counts to determine the most appropriate transcript/polyadenylation site pairs, and 2) compared aptardi-identified polyadenylation sites to annotated polyadenylation sites via the reference transcriptome or the StringTie reconstruction.

Filtering transcript/polyadenylation site pairs based on estimated read count dramatically reduced the number of transcripts a 3' terminus was associated with. The aptardi analysis identified 71,757 new transcript/polyadenylation site pairs (i.e., aptardi transcripts). These represent 34,003 unique 3' termini resulting in each unique 3' terminus being associated with approximately two transcripts on average. After filtering to yield the DABG transcriptome, there were 18,504 aptardi transcripts representing 14,388 3' termini (approximately 1.3 aptardi transcripts associated with each 3' terminus).

Within this pipeline, aptardi interrogates the 3' end of all transcripts. Because of this, it is possible for aptardi for identify a 3' terminus in a transcript/polyadenylation site pair that has been identified using another source (i.e., Ensembl or StringTie) although it has never been associated with the transcript in the aptardi-identified transcript/polyadenylation site pair. Of the initial 71,757 aptardi transcripts, 19% matched the 3' terminus of a reference or StringTie transcript +/- 100 bases, i.e., the 3' terminus was not novel but had not been paired with that particular transcript before. When the percent was calculated based on the number of 3' termini

rather than the number transcripts, a similar percentage was observed (21%; Supplementary Table 4). Of the 18,504 DABG aptardi transcripts, 16% had a 3' terminus that matched annotation (+/- 100 bases) from the Ensembl reference or StringTie transcriptome. This overlap was similarly 16% when calculated based on number of 3' termini (Supplementary Table 4). The slight decrease in the number of aptardi 3' termini matching a StringTie or reference 3' termini in the DABG transcriptome compared to pre-filtering suggests the filtering removes some false positive aptardi transcript/polyadenylation site pairs in favor of the original StringTie/reference transcript with the given polyadenylation site. At the same time, the relatively high number of overlapping reference/StringTie and aptardi 3' termini in the DABG transcriptome suggests that aptardi annotation of the polyadenylation site for the given transcript is accurate and simply represents a transcript with a similar 3' terminus to another StringTie/reference transcript with different upstream exon structure.

*Heritability of Transcripts in the Detected Above Background Transcriptome*

Transcripts in the DABG transcriptome displayed similar heritability regardless of the source (reference, StringTie, or aptardi) used to identify the transcripts (Figure 4.2).

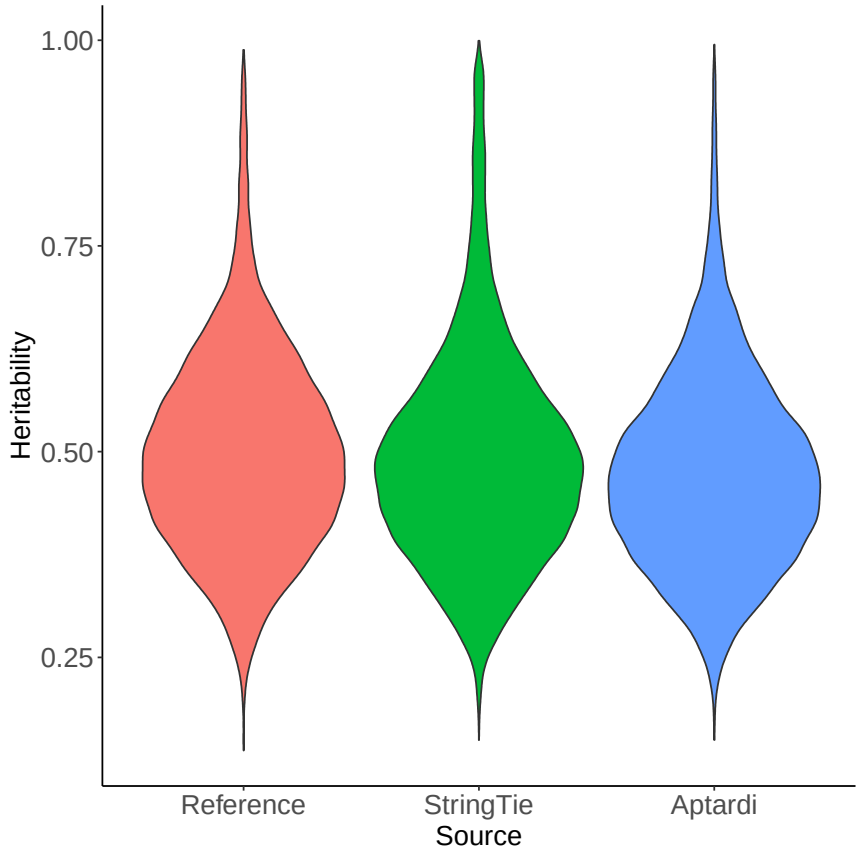



**Figure 4.2. Heritability of transcripts in the detected above background transcriptome.** Heritability of transcripts derived from reference annotation, StringTie, and aptardi. Heritability was estimated as the R-squared value from a one-way ANOVA of individual rat expression values using strain as the predictor (30 strains total) and transcript normalized expression estimates as the response.

**Evaluation of Genes Previously Identified as Associated with Voluntary Alcohol Consumption in the HXB/BXH Recombinant Inbred Rat Panel**

Our earlier work identified a candidate brain coexpression module associated with voluntary alcohol consumption in the HXB/BXH RI panel using RNA expression levels measured from microarrays [132, 445]. As stated previously, this network was generated using data from exon arrays and therefore, lacked the resolution to estimate expression levels of all possible isoforms, i.e., alternative splicing and alternative polyadenylation transcripts. However,

we were able to replicate many of our previous findings for our strongest candidates and gained valuable insight into the transcriptome structure of these candidates.

In particular, we examined the genes from the previous candidate coexpression module that were either significantly associated with alcohol consumption ($p < 0.05$) or were within the top 8 most highly connected genes within the module. The RNA-Seq data from the HXB/BXH RI panel replicated the association with alcohol consumption (correlation in the same direction and p-value < 0.05) for four of the eight genes (Table 4.2). One unannotated gene (*GENE_27603*) from the previous module was not identified in the DABG transcriptome. For *Txnip*, the association with alcohol was suggestive ($p = 0.068$). The two remaining genes (*Cfap91* and *Coq5*) both had two transcripts each and neither transcript was associated with alcohol consumption.

**Table 4.2. Genes whose brain expression were previously identified as associated with voluntary alcohol consumption in the HXB/BXH recombinant inbred rat panel.**

| | | From Microarray Data (Saba et al. 2021) | | From HXB/BXH RNA-Seq Data | | |
|---|---|---|---|---|---|---|
| Gene Symbol | Gene Description | Correlation with Alcohol Consumption with Microarray [Correlation Coefficient (P-Value)] | Connectivity-Based Intramodular Connectivity (Rank Within Module) | # Transcripts Identified in HXB/BXH Panel | # Reference / StringTie / Aptardi Transcripts | Most Significant Transcript Correlation with Alcohol Consumption [Correlation Coefficient (p-value)] |
| Lrap | Locus regulating alcohol preference | -0.55 (0.011) | 2.99 (1) | 1 | 0/1/0 | -0.45 (0.042) |
| Ift81 | Intraflagellar transport 81 | -0.43 (0.051) | 2.66 (2) | 3 | 0/1/2 | -0.44 (0.049) |
| Coq5 | Coenzyme Q5, methyltransferase | -0.50 (0.021) | 2.24 (3) | 2 | 1/0/1 | 0.13 (0.59) |
| Txnip | Thioredoxin interacting protein | 0.61 (0.003) | 2.20 (4) | 1 | 1/0/0 | 0.41 (0.068) |
| P2rx4 | Purinergic receptor P2X 4 | -0.63 (0.002) | 2.15 (5) | 2 | 1/1/0 | -0.58 (0.006) |
| Tmem116 | Transmembrane protein 116 | 0.34 (0.133) | 2.00 (6) | 1 | 1/0/0 | 0.52 (0.017) |
| Cfap91 (formerly Maats1) | Cilia and flagella associated protein 91 | -0.56 (0.008) | 1.95 (7) | 2 | 1/0/1 | 0.21 (0.37) |
| GENE_27603 | Unannotated gene | -0.51 (0.021) | 1.74 (8) | NOT INCLUDED IN DABG TRANSCRIPTOME | | |

Correlations were determined using Spearman's rank correlation and strain mean gene level normalized expression estimates or strain mean transcript level normalized expression estimates and strain mean alcohol consumption values for the gene and transcript correlations, respectively. Genes are ordered by gene correlation with voluntary alcohol consumption p-value. The number of transcripts identified for each gene in the detected above background transcriptome is shown with its source of identification. Maximum transcript correlations with voluntary alcohol consumption are based on absolute values. The intra-modular connectivity values of genes in the candidate module from the previous study (Saba et al. 2021) are shown, along with their rank within the module. Only genes that could be assigned a gene name from the previous study (13 of the 17) were included. All p-values are unadjusted for multiple testing.

Furthermore, our earlier work identified a long, potentially non-coding RNA transcript, *Lrap*, as a key modulator of voluntary alcohol consumption in the HXB/BXH RI panel initially using a systems genetics approach and subsequently using genetically manipulated rats [132, 445]. Since this transcript was originally identified using short read RNA sequencing from the progenitor strains only and does not exist in the reference annotation, we assessed whether it was recapitulated with the full HXB/BXH RI panel through our transcriptome reconstruction methods (i.e., StringTie and/or aptardi). A similar transcript (transcript ID = *MSTRG.6520.1*;

associated gene name = AABR07036336.2) – identified by StringTie – was observed (Figure 4.3). *Lrap* and *MSTRG.6520.1* reside on the negative strand on chromosome 12, possess three exons, and have similar exon structures with near identical intron junctions (Original *Lrap*: 39,009,809-39,016,585, 39,017,055-39,017,223, and 39,021,009-39,021,641; *MSTRG.6520.1*: 39,011,243-39,016,585, 39,017,056-39,017,223, and 39,021,010-39,021,635). Furthermore, *MSTRG.6250.1* was the only isoform of this gene (*MSTRG.6520*). As a result, this gene/transcript was hereafter labeled *Lrap*. With the new RNA-Seq data from the entire HXB/BXH RI panel, *Lrap* remained significantly negatively associated with alcohol consumption (correlation coefficient = -0.45; p-value = 0.042; Table 4.2).

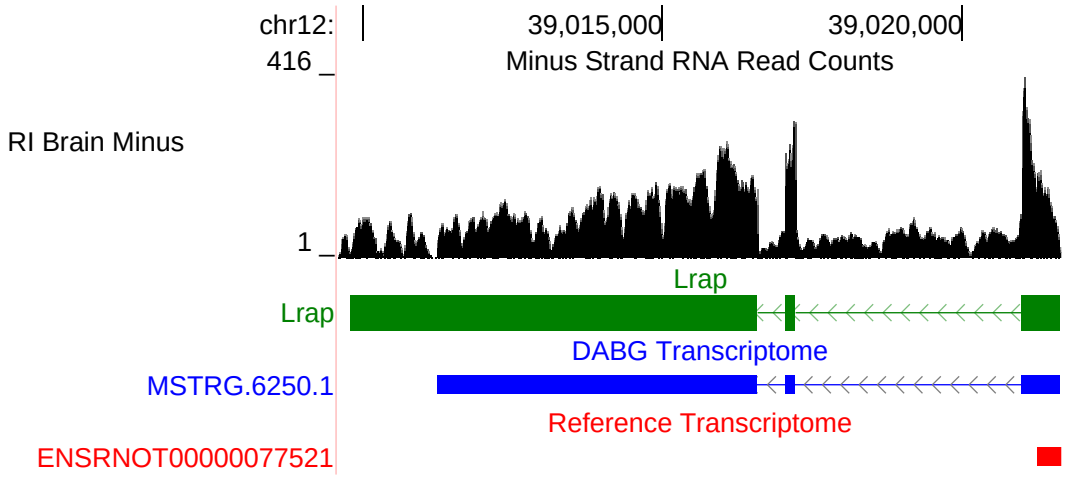



**Figure 4.3. Recapitulation of *Lrap*.** We previously identified a novel transcript, subsequently annotated *Lrap*, as a mediator of alcohol consumption and expressed in rat brain (Saba *et al.* 2015, Saba *et al.* 2021). The structure identified previously is shown in green (*Lrap*), the *de novo* transcript identified in the detected above background (DABG) transcriptome is shown in blue (*MSTRG.6250.1*), and existing reference annotation is shown in red (*ENSRNOT00000077521*). The transcript identified here as *MSTRG.620.1* (annotated by StringTie) closely matches our previous annotation of *Lrap*. Additionally, the RNA sequencing reads on the negative strand (black plot) support the present of this transcript in our dataset. The RNA sequencing reads represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser [381] (http://genome.ucsc.edu).

**Voluntary Alcohol Consumption Quantitative Trait Loci**

Using the 21 HXB/BXH RI strains with voluntary alcohol consumption data and genotype data (Supplementary Table 1) resulted in four suggestive (p-value < 0.63) pQTL; two on chromosome 1, one on chromosome 5 and one on chromosome 12 (Figure 4.4). To deduce if the two suggestive peaks on chromosome 1 represented individual peaks, a second QTL analysis was done that included the maximum peak on chromosome 1 as a covariate (Supplementary Figure 4.4). Since the second QTL did not include any peaks on chromosome 1, the two peaks on chromosome 1 likely represent regions in linkage disequilibrium and were treated as a single QTL in the remainder of the analysis. Notably, the other pQTL on chromosomes 5 and 12 remained suggestive in the second QTL analysis, indicating these pQTL are independent of the pQTL on chromosome 1.

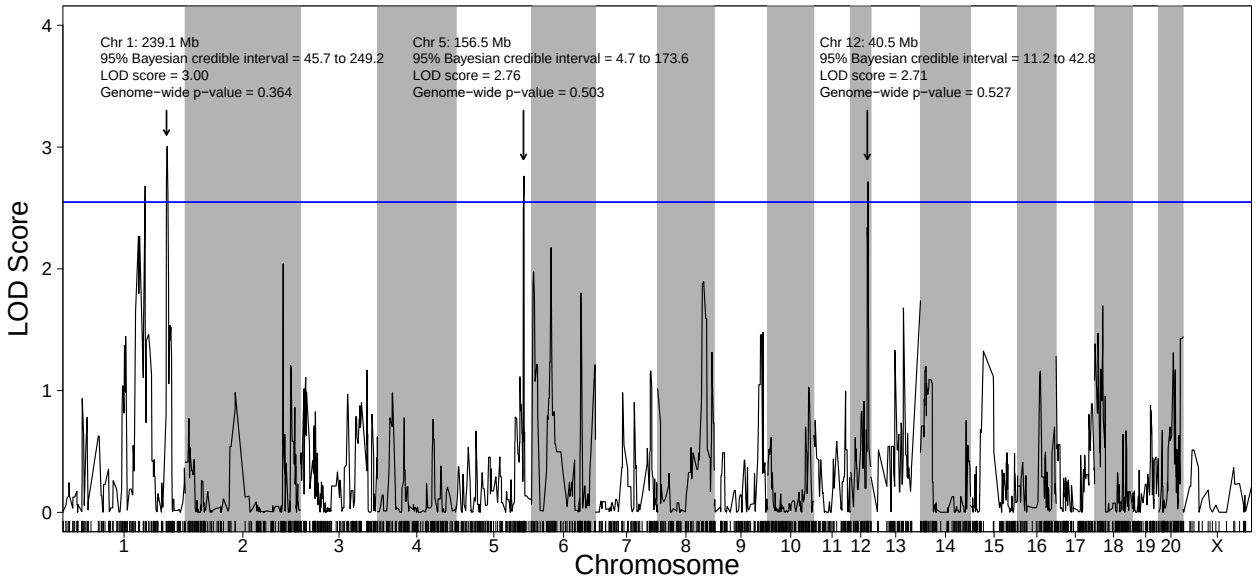



**Figure 4.4. Quantitative trait loci (QTL) for voluntary alcohol consumption in the HXB/BXH recombinant inbred panel.** Strain means were used in a marker regression to determine phenotypic QTL. The blue line represents the logarithm of odds (LOD) score threshold for a suggestive QTL (genome-wide p-value = 0.63). Suggestive QTL are labeled with their location, 95% Bayesian credible interval, LOD score, and genome-wide p-value. Empirical genome-wide phenotypic QTL p-values were calculated using 1,000 permutations.

## Identification of Candidate Coexpression Networks and Candidate Individual Transcripts Associated with Voluntary Alcohol Consumption

RNA expression data from alcohol naïve rats were used to determining candidate networks/ individual candidate transcripts that predispose these animals to voluntary alcohol consumption. Prior to these analyses, additional filtering of the DABG transcriptome was performed to include only transcripts that 1) were the dominant isoforms expressed for a gene (maximum number of transcripts considered per gene = 3) in rats with voluntary alcohol consumption data, 2) demonstrated heritability in the HXB/BXH RI panel and thus genetic influence on RNA expression levels, and 3) could be associated with a gene name (i.e., shared at least one splicing junction with an Ensembl gene) for interpretation purposes. Of the 59,751 transcripts in the DABG transcriptome, 37,453 were kept after removing transcripts that were not within the top three expressed isoforms for a gene. Eliminating transcripts not highly heritable (heritability <= 0.478), resulted in 20,442 transcripts. Finally, removing transcripts without an associated gene symbol produced a final set of 18,543 transcripts (Supplementary Figure 5). The final set of transcripts were derived from 12,609 genes, of which 7,945 (63%), 3,403 (27%), and 1,261 (10%) possessed one, two, and three isoforms, respectively. Of the 18,543 transcripts, 5,427, 4,932, and 8,175 transcripts were identified by aptardi, StringTie, and the reference, respectively.

*Candidate Individual Transcripts*

Of the 18,534 individual transcripts whose strain mean normalized expression estimates were subjected to correlation analysis with strain mean voluntary alcohol consumption, 64 were significantly (p-value < 0.01) correlated with voluntary alcohol consumption. Requiring a significant (genome-wide p-value < 0.01) eQTL, as well as eQTL overlap with a pQTL (using

95% Bayesian credible intervals) resulted in a final set of 11 transcripts (Table 4.3). One of these transcripts was identified by aptardi, six were identified by StringTie, and four were identified by the reference. Furthermore, seven of the 11 transcripts belonged to genes expressing multiple isoforms. Other notable transcripts include *Map3k7*, which possesses a distal eQTL that overlaps the pQTL for alcohol consumption on chromosome 5, and *Aldh1a7*, which has a clear connection to alcohol. Besides Map3k7, all transcripts contained local eQTL. We note that using a multiple testing correction p-value eliminated all transcripts – likely due to the small sample size – but suggest applying additional filtering criteria reduces the number of false positives.

**Table 4.3. Individual candidate transcripts in brain for predisposition to voluntary alcohol consumption.**

| Transcript ID | Gene Symbol(s) | Gene Description(s) | Source | Transcript Correlation With Alcohol Consumption [Correlation Coefficient (P-Value)] | Expression QTL LOD Score [Genome-Wide (P-Value)] | Expression QTL Chromosome:Position (Mb) | # Transcripts Identified in HXB/BXH RI Panel | # Reference / StringTie / Aptardi Transcripts |
|---|---|---|---|---|---|---|---|---|
| ENSRNOT00000072618 | *E2f2* | E2F transcription factor | Reference | -0.66 (0.0013) | 13.88 (<0.001) | 5:154.8 | 1 | 1/0/0 |
| MSTRG.1868.13 | *Tmem9b* | Transmembrane protein 9b | StringTie | -0.62 (0.0030) | 5.26 (<0.001) | 1:173.9 | 11 | 0/10/1 |
| MSTRG.1793.4 | *Trim68* | Tripartite motif-containing 68 | StringTie | 0.61 (0.0031) | 13.06 (<0.001) | 1:167.2 | 4 | 3/1/0 |
| ENSRNOT00000075003 | *Tmem159* | Transmembrane protein 159 | Reference | -0.60 (0.0038) | 9.33 (<0.001) | 1:189.2 | 1 | 1/0/0 |
| MSTRG.23809.1 | *Map3k7* | Mitogen activated protein kinase kinase kinase 7 | StringTie | 0.60 (0.0041) | 5.92 (0.0040) | 5:46.8 | 2 | 1/1/0 |
| ENSRNOT00000090867 | *Oas3* | 2'-5'-oligoadenylate synthase 3 | Reference | 0.60 (0.0043) | 9.29 (<0.001) | 12:40.5 | 1 | 1/0/0 |
| ENSRNOT00000024093.1 | *Aldh1a7* | Aldehyde dehydrogenase, cytosolic 1 | Aptardi | -0.59 (0.0051) | 14.28 (<0.001) | 1:237.6 | 1 | 0/0/1 |
| ENSRNOT00000001752 | *P2rx4* | Purinergic receptor P2X 4 | Reference | -0.58 (0.0059) | 9.81 (<0.001) | 12:39.1 | 2 | 1/1/0 |
| MSTRG.1874.1 | *Tmem41b* | Transmembrane protein 41B | StringTie | 0.57 (0.0068) | 9.37 (<0.001) | 1:173.9 | 3 | 1/2/0 |
| MSTRG.2084.2 | *Lat,Spns1,Nfatc2ip* | Linker for activation of T-cells family member 1, Protein spinster homolog 1, NFATC2-interacting protein | StringTie | 0.56 (0.0089) | 7.65 (<0.001) | 1:197.0 | 7 | 1/6/0 |
| MSTRG.1526.1 | *Pex11a* | Peroxisomal membrane protein 11A | StringTie | 0.55 (0.0093) | 11.02 (<0.001) | 1:141.0 | 2 | 1/1/0 |

Strain mean normalized expression estimates and strain mean voluntary alcohol consumption values were used to determine Spearman's rank correlations. Strain mean normalized expression estimates were used in a marker regression to determine expression quantitative trait loci (eQTL) and corresponding logarithm of odds (LOD) scores. Empirical genome-wide expression QTL p-values were calculated using 1,000 permutations. Transcripts are ordered by p-value. The total number of transcripts generated from the same gene in the detected above background transcriptome, and the source(s) identifying the transcripts, is also shown. All transcripts possessed local eQTL with the exception of the transcript from the *Map3k7* gene.

Map3k7

Two isoforms of Map3k7 (gene ID = *MSTRG.23809*) were present in the DABG transcriptome (Figure 4.5A), of which one was from the reference (*ENSRNOT00000007657*) and one was identified by StringTie *(MSTRG.23809.1*). *MSTRG.23809.1*, the candidate individual transcript, is located on the plus strand of chromosome 9 (114.02-114.07 Mb), but its eQTL

overlapped the voluntary alcohol consumption pQTL on chromosome 5. The transcript structure of *MSTRG.23809.1* represents an exon skipping isoform of *ENSRNOT00000007657*; specifically, it lacks exon 12 (Figure 4.5B). Across the 63 individual rat RNA-Seq libraries (21 HXB/BXH RI strains) with voluntary alcohol consumption data, the mean TPM of *MSTRG.23809.1* and *ENSRNOT00000007657* was 0.95 and 1.30, respectively.

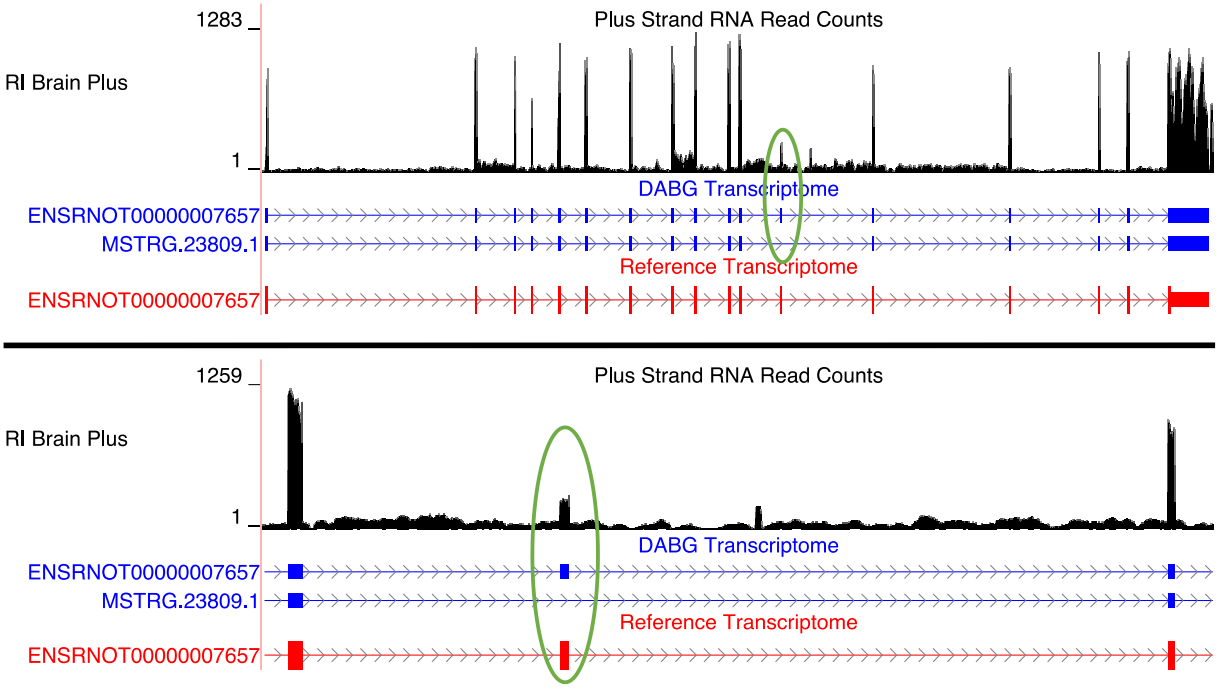



**Figure 4.5. Isoforms of the *Map3k7* gene. (A)** Blue transcripts represent those identified in the detected above background (DABG) transcriptome, and the red transcript represents the transcript present in reference annotation. *ENSRNOT00000007657* is annotated in the reference transcriptome and retained in the DABG transcriptome. *MSTRG.23809.1* represents a novel isoform identified by StringTie. *MSTRG.23809.1* represents an isoform with exon 12 skipped compared to *ENSRNOT00000007657* (circled in green). **(B)** A zoomed in image of the exon 12 region. The RNA sequencing reads on the positive strand (black plot) represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser [381] (http://genome.ucsc.edu).

We note that a transcript with an identical transcript structure as *MSTRG.23809.1* was identified *de novo* by StringTie (*MSTRG.1784.2*) on the plus strand of chromosome 5 (47.19-47.24 Mb), which overlaps the voluntary alcohol consumption pQTL. The gene of this transcript (*LOC100910771*) was previously described as *Map3k7-like* by the Rat Genome Database, but

has since been re-annotated as *Map3k7*. Therefore, there are multiple locations of this gene/transcript currently annotated in rat. Moreover, the strain mean normalized expression estimates of *MSTRG.23809.1* and *MSTRG.1784.2* are highly correlated (Spearman's rank correlation = -0.84). While both transcripts' strain mean normalized expression estimates have a similar (absolute) Spearman's rank correlation to strain mean voluntary alcohol consumption (*MSTRG.23809.1* = 0.600; *MSTRG.1784.2* = -0.509), the slightly weaker correlation of *MSTRG.1784.2* caused it to be removed as a candidate transcript based on correlation p-value (p-value = 0.018). Taken together, we suggest the transcript structure is accurate (i.e., the exon 12 skipping isoform) and, furthermore, the transcript is likely a bona fide candidate individual transcript, but note the genomic location of the candidate individual transcript is unclear.

Aldh1a7

A single transcript of *Aldh1a7*, *ENSRNOT00000024093.1*, was present in the DABG transcriptome (Figure 4.6). The reference version of this transcript, *ENSRNOT00000024093*, was removed during filtering and thus not present in the DABG transcriptome. *ENSRNOT00000024093.1* shares intron junctions with *ENSRNOT00000024093* (Figure 4.6A) but possesses a unique 3' terminus on the negative strand of chromosome 1 compared to the reference (*ENSRNOT00000024093.1* 3' end = 240,561,896; *ENSRNOT00000024093* 3' end = 240,562,423; Figure 4.6B).

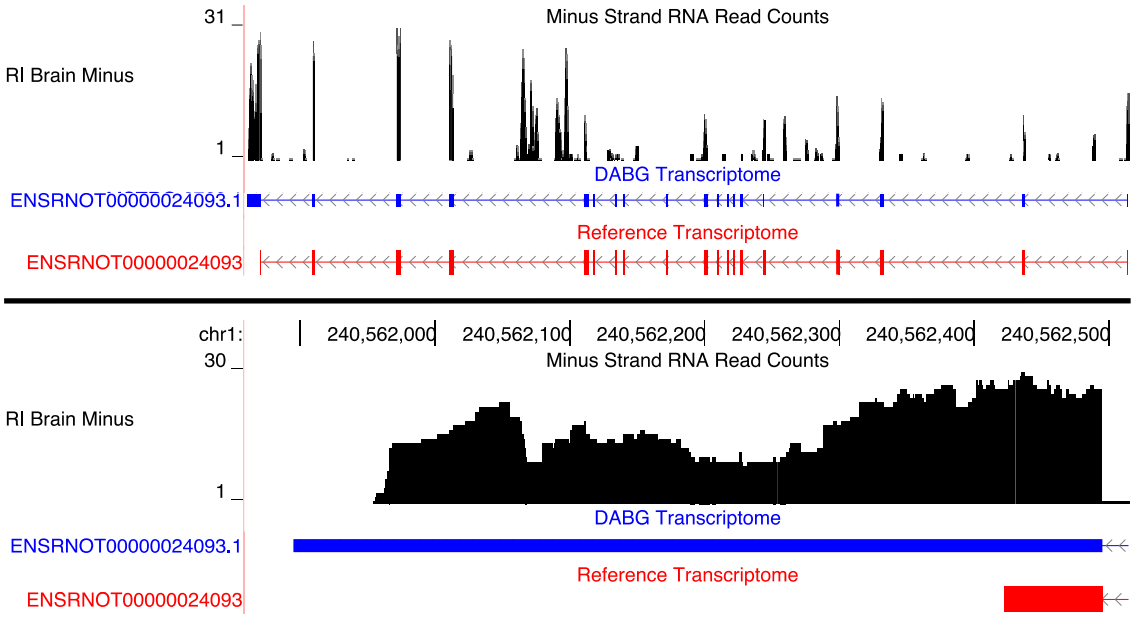



**Figure 4.6. Isoforms of the *Aldh1a7* gene. (A)** The blue transcript represents those identified in the detected above background (DABG) transcriptome, and the red transcript represents the transcript present in reference annotation. *ENSRNOT00000024093* is annotated in the reference transcriptome but was filtered out of the DABG transcriptome. *ENSRNOT00000024093.1* represents a novel isoform identified by aptardi. (**B)** A zoomed in image of the 3' region comparing aptardi annotation (*ENSRNOT00000024093.1*) to reference annotation (*ENSRNOT00000024093*). *ENSRNOT00000024093* and *ENSRNOT00000024093.1* differ only in the length of their 3' most exon. The RNA sequencing reads on the negative strand (black plot) represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser [381] (http://genome.ucsc.edu).

*Candidate Modules from Weighted Gene Coexpression Network Analysis*

A total of 30 HXB/BXH RI strains with expression data (Supplementary Table 1) were used to generate transcript coexpression modules using strain means of transcript normalized expression estimates. WGCNA identified 215 modules along with 137 transcripts (out of the 18,543) that were not assigned a module. The median module size was 10 transcripts (Supplementary Figure 6). Module eigengenes captured much of the within-module transcript expression variability (interquartile range: 60% to 72%). In addition, many isoforms of genes belonged to different modules (Supplementary Figure 7). For example, of the 4,664 genes with

more than one isoform included in WGCNA, 2,526 (54%) genes possessed isoforms that belonged to more than one module.

Of the 215 modules, the module eigengene of a single module – blue1 – was significantly (p-value < 0.01) associated with voluntary alcohol consumption (correlation coefficient = -0.62, p-value = 0.0026). Using the 30 HXB/BXH RI strains with blue1 module eigengene values and genotype data (Supplementary Table 1), a genome-wide significant (p-value < 0.01) meQTL was identified on chromosome 12 (logarithm of odds score = 16.83, p-value < 0.0001). Furthermore, the location of the meQTL (chromosome 12, position = 39.1 Mb, 95% Bayesian credible = 39.1-40.5 Mb) overlapped the suggestive (p-value < 0.63) pQTL on chromosome 12 (chromosome 12, position = 40.5 Mb, 95% Bayesian credible = 11.2-42.8 Mb) thereby satisfying all the requirements for candidacy. The module eigengene explained 75% of the within-module expression variability.

The transcripts comprising the blue1 module are shown in Figure 4.7 and listed in Table 4.4. One transcript was identified by aptardi (*Ift81*), three were identified by StringTie (*Lrap*, *Mapkapk5*, *AABR07065438.1*) and two were in reference annotation (*P2rx4* and *Oas3*). Most of these transcripts reside near the physical location of the meQTL and pQTL. Three of the six transcripts shared a gene ID with other transcripts (*Mapkapk5*, *P2rx4*, and *Ift81*) in the DABG transcriptome, indicating these transcripts represented splicing variants (i.e., isoforms) of the gene. The expression of all transcripts displayed individual correlation with voluntary alcohol consumption (p-value < 0.05). The genes of these three transcripts (*Lrap*, *P2rx4* and *Ift81*) were identified in our previous candidate module [132] using microarray data (vs RNA-Seq data here).

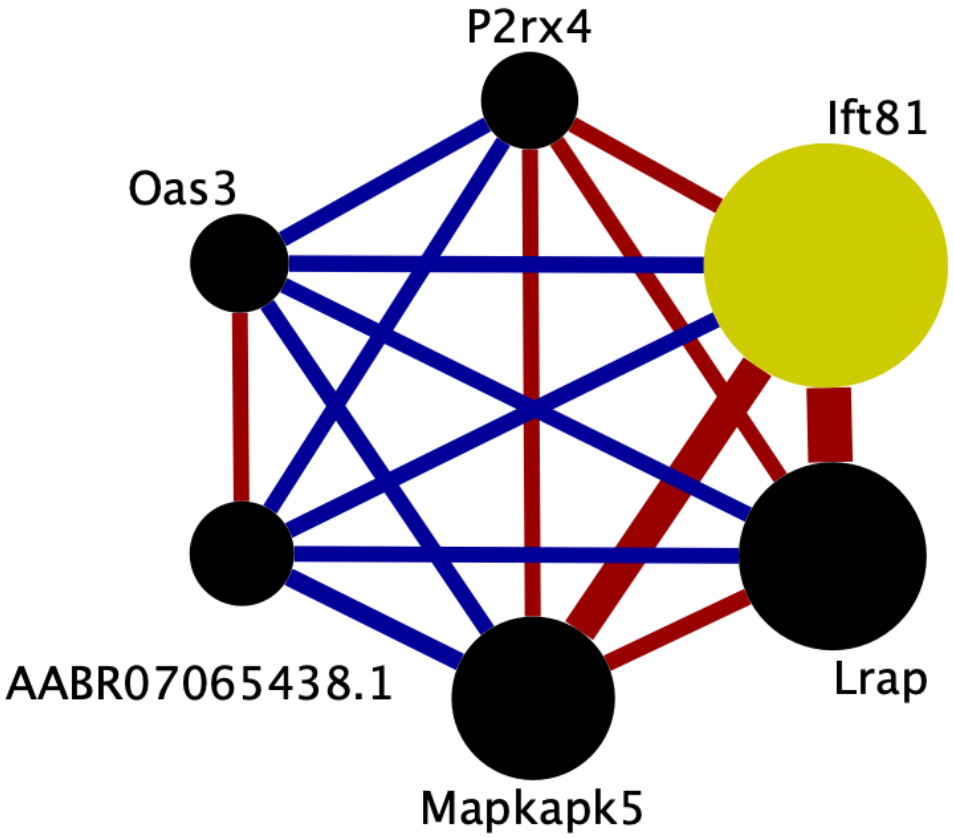



**Figure 4.7. Connectivity within the brain candidate coexpression module for predisposition to voluntary alcohol consumption.** Each circle represents a transcript from the coexpression module. The size of each circle is weighted based on its intra-modular connectivity (not to scale), and the thickness of each edge is weighted based on the magnitude of the connectivity between the two transcripts (not to scale). The edge colors indicate the direction of the connectivity (red = positive, blue = negative). The hub transcript, defined here as the single transcript with the largest intra-modular connectivity, is colored in yellow (Ift81), and its expression is negatively associated with voluntary alcohol consumption. Transcripts are ordered by intra-modular connectivity clockwise starting with Ift81. Strain mean normalized transcript expression estimates were used in weighted gene coexpression network analysis to generate coexpression networks. This figure was generated using Cytoscape (v. 3.8.2) [446].

**Table 4.4. Transcripts in the brain candidate coexpression module for predisposition to voluntary alcohol consumption.**

| Transcript ID | Gene Symbol | Gene Description | Source | Chromosome:Start Position-End Position (Mb) (Strand) | Intra-modular Connectivity | Transcript Correlation With Alcohol Consumption [Correlation Coefficient (P-Value)] | # Transcripts Identified in HXB/BXH RI Panel | # Reference / StringTie / Aptardi Transcripts |
|---|---|---|---|---|---|---|---|---|
| ENSRNOT00000066952.1 | Ift81 | Intraflagellar transport 81 | Aptardi | 12:39.42-39.51 (+) | 1.22 | -0.44 (0.0487) | 3 | 0/1/2 |
| MSTRG.6250.1 | Lrap | Locus regulating alcohol preference | StringTie | 12:39.01-39.02 (-) | 1.14 | -0.45 (0.0417) | 1 | 0/1/0 |
| MSTRG.6281.1 | Mapkapk5 | MAPK activated protein kinase 5 | StringTie | 12:40.51-40.53 (+) | 1.10 | -0.49 (0.0238) | 3 | 2/1/0 |
| MSTRG.19929.1 | AABR07065438.1 | Ribosomal protein L6, pseudo 1 | StringTie | 6:128.74-128.74 (+) | 1.01 | 0.51 (0.0191) | 1 | 0/1/0 |
| ENSRNOT00000090867 | Oas3 | 2'-5'-oligoadenylate synthetase 3 | Reference | 12:41.32-41.34 (+) | 1.01 | 0.60 (0.0043) | 1 | 1/0/0 |
| ENSRNOT00000001752 | P2rx4 | Purinergic receptor P2X 4 | Reference | 12:39.31-39.33 (-) | 1.00 | -0.58 (0.0059) | 2 | 1/1/0 |

Transcripts are ordered by intra-modular connectivity, and the source that identified the transcript is shown. The total number of transcripts with the same gene ID as the candidate individual transcripts (i.e., isoforms of the gene) in the detected above background transcriptome, and the source(s) identifying the transcripts, is also shown. Strain mean normalized expression estimates and strain mean voluntary alcohol consumption values were used to determine Spearman's rank correlations. Strain mean normalized transcript expression estimates were used in weighted gene coexpression network analysis to generate coexpression networks.

P2rx4

Two isoforms of *P2rx4* (gene ID = *MSTRG.6256*) were identified in the DABG transcriptome: *ENSRNOT00000001752*, which was annotated in the reference transcriptome and represented the transcript in the candidate module, and *MSTRG.6256.1*, which was annotated by StringTie (Figure 4.8). The reference and StringTie transcripts differ at their 5' most exon but otherwise share identical transcript structure. Both were included in WGCNA.

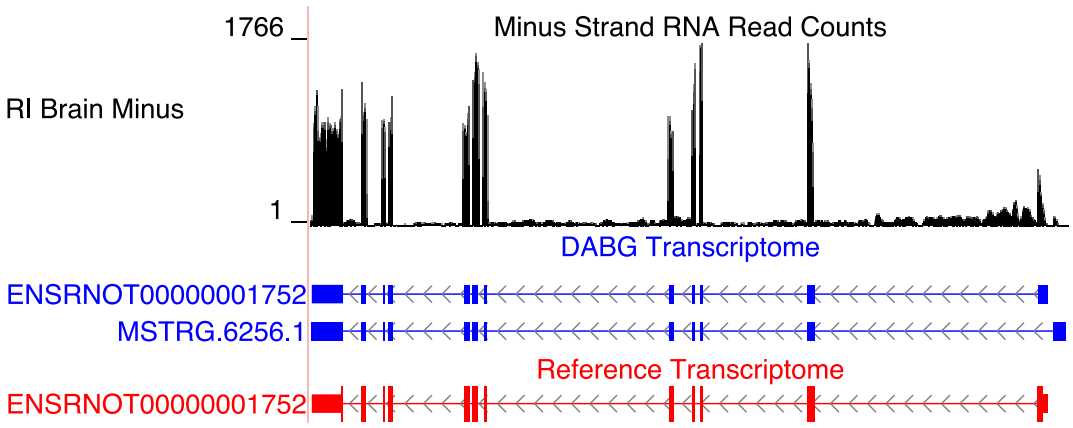



**Figure 4.8. Isoforms of the *P2rx4* gene.** Blue transcripts represent those identified in the detected above background (DABG) transcriptome, and the red transcript represents the transcript present in reference annotation. *ENSRNOT00000001752* is annotated in the reference transcriptome and retained in the DABG transcriptome, whereas *MSTRG.6256.1* represents a novel isoform identified here by StringTie. *MSTRG.6256.1* and *ENSRNOT00000001752* differ in their 5' most exon but otherwise share exon structure. The RNA sequencing reads on the negative strand (black plot) represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser [381] (http://genome.ucsc.edu).

*ENSRNOT00000001752* was also identified as a candidate transcript (see Individual Candidate Transcripts in Results). While the strain mean normalized expression estimates of *ENSRNOT00000000175* were negatively correlated with strain mean voluntary alcohol consumption (Spearman's rank correlation = -0.579, p-value = 0.0051), *MSTRG.6256.1* was not significantly associated with alcohol consumption (Spearman's rank correlation = 0.166, p-value = 0.471). *ENSRNOT00000001752* was the dominant isoform in the 63 individual rat RNA-Seq libraries (21 HXB/BXH RI strains) with voluntary alcohol consumption data (*ENSRNOT00000001752* mean TPM = 6.143; *MSTRG.6256.1* TPM = 0.254).

Ift81

An isoform of *Ift81*, *ENSRNOT00000066952.1*, is the hub gene in the candidate module (i.e., transcript with the greatest intra-modular connectivity). This transcript was identified by aptardi. Two additional isoforms of this gene were annotated in the DABG transcriptome:

*ENSRNOT00000066952.2*, which was annotated by aptardi, and *MSTRG.6258.1*, which was identified by StringTie (Figure 4.9). The reference annotation of *Ift81*, *ENSRNOT00000066952*, was not present in the DABG transcriptome. Both aptardi transcripts only differ in their 3' base position (on the plus strand of chromosome 12) compared to the reference transcript (*ENSRNOT00000066952* 3' end = 39,506,890; *ENSRNOT00000066952.1* 3' end = 39,507,407; *ENSRNOT00000066952.2* 3' end = 39,507,807; Figure 4.9B). In contrast, the transcript identified by StringTie, *MSTRG.6258.1*, possesses unique exon structure (Figure 4.9A). The isoform belonging to this candidate module (*ENSRNOT00000066952.1*) was the only individual isoform of this gene significantly (p-value < 0.05) associated with voluntary alcohol consumption (*ENSRNOT00000066952.1*: correlation = -0.44, p-value = 0.049; *ENSRNOT00000066952.2*: correlation = -0.33, p-value= 0.14, *MSTRG.6258.1*: correlation = -0.15, p-value = 0.52). Across the 63 RNA-Seq samples of the 21 HXB/BXH RI strains with voluntary alcohol consumption data, the mean transcripts per million was 0.68, 2.35, and 2.67 for *ENSRNOT00000066952.2*, *ENSRNOT00000066952.1*, and *MSTRG.6258.1*, respectively. Only *ENSRNOT00000066952.1*, and *MSTRG.6258.1* were subjected to WGCNA.

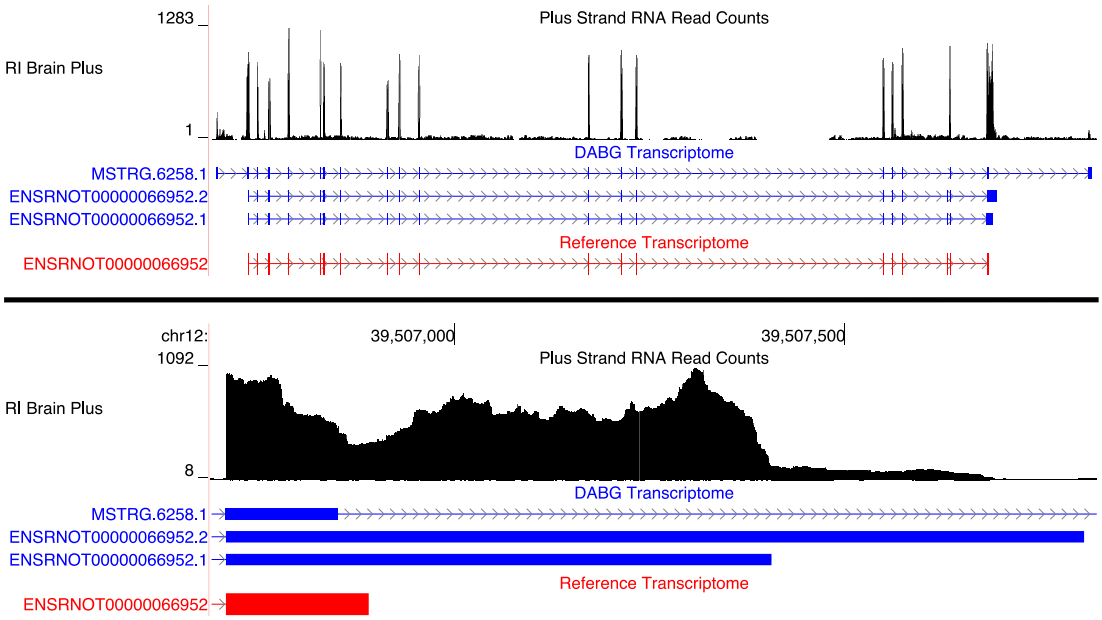



**Figure 4.9. Isoforms of the *Ift81* gene. (A)** Blue transcripts represent those identified in the detected above background (DABG) transcriptome, and the red transcript represents the transcript present in reference annotation. *ENSRNOT00000066952* is annotated in the reference transcriptome but removed in the DABG transcriptome. *MSTRG.6258.1* represents a novel isoform identified by StringTie, and *ENSRNOT00000066952.1* and *ENSRNOT00000066952.2* represent novel isoforms identified by aptardi. (**B)** A zoomed in image of the 3' region comparing aptardi annotation (*ENSRNOT00000066952.1* and *ENSRNOT00000066952.2*) to reference annotation (*ENSRNOT00000066952*). *ENSRNOT00000066952*, *ENSRNOT00000066952.1*, and *ENSRNOT00000066952.2* differ only in the length of their 3' most exon, whereas *MSTRG.6258.1* possesses a different exon structure. The RNA sequencing reads on the positive strand (black plot) represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser [381] (http://genome.ucsc.edu).

## Discussion

### Characterization of the Brain Specific Transcriptome in the HXB/BXH Recombinant Inbred Rat Panel

A major goal of this work was to annotate the brain specific transcriptome in the HXB/BXH recombinant inbred rat panel by applying computational methods – namely StringTie and aptardi – that incorporate expression data (in the form of RNA-Seq). By including expression data, we were able to characterize the expressed transcriptome in the specific context

of our study. Moreover, our deeply sequenced RNA-Seq libraries enabled high resolution mapping of the transcriptome.

Many transcripts in the DABG transcriptome were identified by the computational approaches that utilized RNA-Seq data (Table 4.1) and were not included in the current Ensembl transcriptome for rat, highlighting the importance of generating transcriptomes that take into account study-specific expression. Another – likely complementary – explanation of the abundance of StringTie and aptardi transcripts is that the rat reference transcriptome is under annotated compared to humans and other model species such as mouse [447]. Additionally, the comparable expression heritabilities of StringTie and aptardi transcripts to reference transcripts indicates these algorithms annotate genetically meaningful transcripts.

Unsurprisingly, we found that the transcript:gene ratio in the DABG transcriptome was much greater than existing rat Ensembl reference annotation (Table 4.1), and many more genes in the DABG transcriptome expressed more than one isoform compared to the reference (Supplementary Figure 3). This aligns with the literature that many higher order eukaryotic genes express alternative splicing and/or alternative polyadenylation transcripts [102, 412] and, in particular, that brain expresses the greatest mRNA diversity compared to other tissues due to alternative splicing and alternative polyadenylation [448]. Furthermore, a recent study likewise observed an increase in the transcript:gene ratio in rat when including RNA-Seq data to generate the transcriptome compared to reference annotation [447]. The transcripts of multiple isoform genes were mostly identified through multiple sources (i.e., StringTie, aptardi, and reference), demonstrating the utility of our transcript generation pipeline.

Also of note, we suggest that our transcriptome generation pipeline, including the filtering procedure, provides a means to generate a high quality, representative transcriptome.

Interestingly, the filtering procedure resulted in a marked reduction in the number of reference transcripts in the DABG transcriptome. We hypothesize this is because many isoforms identified by StringTie and/or aptardi more accurately annotated the transcripts expressed by the genes here. Specific examples of this include *Ift81* and *Aldh1a7*, where the reference transcript was removed after filtering in favor of the StringTie and/or aptardi transcripts, and these computationally identified transcripts are better supported by the RNA-Seq data (Figure 4.9 and Figure 4.6).

**Evaluation of Genes Previously Identified as Associated With Voluntary Alcohol Consumption in the HXB/BXH Recombinant Inbred Rat Panel**

We previously performed WGCNA on brain expression data in the HXB/BXH RI panel to identify networks of genes associated with the same phenotype (voluntary alcohol consumption) [132]. The major difference here was in the technology used to generate expression data and, as a result, the type of analysis performed. Specifically, we utilized RNA-Seq data here (as opposed to microarray data) for the quantitative measurement of transcript expression. In the previous analysis, we did do a transcriptome reconstruction using RNA-Seq data from the HXB/BXH progenitor strains but we had to rely on the Affymetrix array to have probes that distinguished between transcripts of a gene. For most genes, the array was not capable of unambiguously estimating the expression of individual transcripts. With RNA-Seq data from the full panel, we were not only able to estimate expression for each individual transcript of a gene, but we were also able to filter individual transcripts for their heritability. As a result, we sought to compare the results of this analysis compared to our earlier work.

*Comparison of the Candidate Transcript Coexpression Network Identified Here to the Previous Candidate Gene Coexpression Network*

The previous candidate coexpression module's meQTL overlapped a voluntary alcohol consumption pQTL on chromosome 12. Likewise, here we identified a single candidate coexpression module (out of 215 modules) with its meQTL/pQTL overlap also on chromosome 12. Of the six transcripts in the new candidate transcript module, genes of three were present in our previous candidate gene module – *P2rx4*, *Lrap*, and *Ift81*. Additionally, the genes of these transcripts had some of the greatest intra-modular connectivity values in the previous gene module; *Lrap*, *Ift81*, and *P2rx4* had the first, second, and fifth greatest intra-modular connectivity values, respectively. In particular, *Lrap* was identified as a key modulator of voluntary alcohol consumption in these rats [445]. For many other genes from the original candidate coexpression module – especially those with high intra-modular connectivity values – displayed correlation with voluntary alcohol consumption at the individual transcript level (Table 4.2). Overall, these results indicate that the major genes/transcripts in candidate coexpression modules (i.e., those with the greatest intra-modular connectivity values) are robust across diverse data and experimental designs.

*Recapitulation of Lrap*

The hub gene of the previous candidate coexpression module was an unannotated gene initially identified from the transcriptome reconstruction in the progenitor strains. Its structure was confirmed using PCR and was subsequently annotated it as long non-coding RNA for alcohol preference, *Lrap*. In this study, we were able to *de novo* assemble its transcript structure (Figure 4.3) using all of the HXB/BXH RI strains, thereby providing additional experimental validation of our computationally identified transcript. The intron junctions of these two

transcripts are nearly identical; however, the length of the 3' end location differs to a greater degree (39,009,809 here vs 39,011,243 previously). We note that the precise locations of the 3' and 5' ends have never been validated [132].

**Candidate Coexpression Network and Candidate Individual Transcripts**

*Characterization of Transcripts in the Candidate Coexpression Network*

Existing reference annotation possessed a single transcript structure for *Ift81*; here we identified three alternative transcripts (Figure 4.9). Moreover, filtering to yield the DABG transcriptome removed the reference transcript, but the three new structures remained. The RNA-Seq data support the presence of these structures, exemplifying how the transcriptome generation procedure can not only annotate new transcripts, but also potentially improve annotation of existing transcripts. Furthermore, while *Ift81* was identified in the previous candidate gene module and was present in the candidate transcript module, only a single isoform of *Ift81* was present in the candidate transcript module, thereby enabling greater granularity as to the exact transcript that is associated with alcohol consumption. Likewise, a second, previously unannotated isoform of *P2rx4* was identified, but the reference transcript of this gene was present in the candidate transcript module. Of note, both of these transcripts were the predominantly expressed isoforms for their respective genes.

Beyond *Ift81*, *P2rx4*, and *Lrap*, the other transcripts in the candidate module include isoforms/transcripts of *Mapkapk5*, *Oas3*, and *AABR07065438.1* (an unannotated transcript).

The *Mapkapk5*, MAPK activated protein kinase 5, transcript in the candidate module was annotated by StringTie (*MSTRG.6281.1*). There are two additional isoforms of *Mapkapk5* in the DABG transcriptome which were annotated by the reference (*ENSRNOT00000001817* and *ENSRNOT00000065314*); however, only *MSTRG.6281.1* was included in WGCNA.

*MSTRG.6281.1* shares similar exons to the longer reference isoform *ENSRNOT00000001817* but with a noticeably longer 3' terminal exon and an additional, long 5' exon (Supplementary Figure 8). MAPK activated protein kinases are enzymes whose activation is mediated by Mapks [449]. Notably, Mapkapk5 is a downstream target of Mapk14 [450], which was shown to be a central regulator of the immunological response in astrocytes [451]. *Mapk14* was also differentially expressed in alcohol preferring AA (alko, alcohol) vs alcohol-avoiding (alko, non-alcohol) rats [452, 453]. Furthermore, *Mapk14* was expressed at lower levels in alcohol preferring iP rats compared to the alcohol non preferring iNP rats in the caudate-putamen of brain [454]. Interestingly, Map3k7 – a candidate individual transcript – is an upstream regulator of Mapk14 [455], providing a common biological pathway for *Mapkapk5* and *Map3k7*.

The single transcript for *Oas3*, 2'-5'-oligoadenylate synthetase 3, in the candidate module was annotated by the reference. A genome wide association study of alcohol consumption in Korean male drinkers identified a SNP in *OAS3* with genome wide significance [456]. Similar to Mapkapk5, OAS3 plays a role in immunity, namely the antiviral immune response [457]. Expression of Oas3 was shown to be enriched in infiltrating macrophages relative to homeostatic microglia during virus-induced neuroinflammation [458].

The unannotated transcript is from the *AABR07065438.1* gene and was identified by StringTie. The single reference transcript for this gene was removed from the DABG transcriptome and, therefore, was also not included in WGCNA. The reference and StringTie transcripts have identical 5' and 3' ends but differ in that the reference transcript has as a single exon, whereas StringTie identified a splice junction and therefore annotated two exons (Supplementary Figure 9). Ensembl describes the reference transcript as ribosomal protein L6,

pseudo 1. Pseudogenes have similar sequence to another gene but are defective [459]; however, pseudogenes are expressed [460].

Overall, the candidate module is linked to inflammation/the immune response. Such an observation is consistent with our previous findings [132, 445].

**Candidate Individual Transcripts**

Notable candidate transcripts include *Aldh1a7* and *Map3k7*. The candidate individual transcript of *Map3k7* was identified *de novo* and differed from the existing reference transcript for this gene in that exon 12 was skipped (Figure 4.5). Previous literature has reported that *Map3k7* expresses an exon 12 skipping isoform of the gene [461, 462], providing credence for the transcript structure identified here. Specifically, the exon skipping isoform was observed to be differentially expressed in the JSL1 human T-cell line when stimulated to illicit an immune response [461]. In regard to alcohol, differences in brain expression of *Map3k7* between high and low alcohol preferring mice have been reported. Moreover, Map3k7 pathways were displayed changes in brain expression of adolescent alcohol-preferring rats following binge-like-alcohol drinking [463]. Taken together, these results may indicate that differences in baseline expression levels of the exon 12 skipping Map3k7 isoform may predispose an individual to differences in voluntary alcohol consumption via modulation of immune response.

The humans *ALDH1* family consists of six genes [464]. Aldh1a7 is an additional rodent-specific gene for this family [465] that is a paralogue of Aldh1a1 [56]. Here we annotated a novel transcript for this gene, *ENSRNOT00000024093.1*, which was identified by aptardi. This transcript was present in the DABG transcriptome, while the reference transcript for this gene (*ENSRNOT00000024093*), was removed during filtering. The transcripts shared intron junctions but differ in that the aptardi transcript has a longer 3’ exon. The RNA-Seq reads support the

presence of the aptardi transcript. Aldh1a7 catalyzes the irreversible conversion of retinaldehyde to retinoic acid [466]. Retinoic acid can act on immune cells and is involved in neuroinflammation [467].

These candidate individual transcripts further point towards the role of inflammation and immunity as predisposing factors associated with voluntary alcohol consumption.

# CHAPTER V

# SUMMARY AND FUTURE DIRECTIONS

## Summary

### Aim 1: Develop a systems genetics pipeline for identifying networks of genes associated with a complex trait (Chapter II).

In Chapter II, we sought to demonstrate/validate a pipeline that takes a systems genetics approach for understanding the genetic architecture of complex traits. Prior work has utilized quantitative genetics analysis to integrate physiological, behavioral, and transcriptomic information to uncover a number of genetic factors for predisposition to cardiovascular, metabolic, and certain behavioral traits [468, 469]. We have adopted this approach and integrated several filters to focus attention on the role of coexpression modules, and we have required conditions that have to be met to categorize a module as a candidate for influencing the quantitative character of a chosen trait. In the current work we chose to apply our approach to analysis of two phenotypes related to alcohol (ethanol) metabolism: 1) alcohol clearance and 2) the measure of circulating acetate levels over time after alcohol administration (i.e., the "area under the curve" for acetate). With regard to the alcohol clearance phenotype, our hypothesis was that if our approach was viable, the identified module would contain components, such as alcohol dehydrogenases, which are accepted determinants of the rate of alcohol metabolism in mammals. For the phenotype of acetate "area under the curve", the approach was being used as hypothesis-generating rather than a hypothesis testing entity.

Overall, the unsupervised, statistically based, systems biology approach that we instituted for analyzing factors influencing ethanol metabolism and resultant acetate levels produced some rewarding results. First, out of 658 modules, our approach identified one module related to the

genomic locus determining the rate of ethanol clearance. This liver module contained two alcohol dehydrogenase transcripts that would be fully expected, from ample literature [278, 279], to be responsible for ethanol oxidation in the rat. The identification of a module with two alcohol dehydrogenases also substantiates the belief that alcohol dehydrogenase isoforms with different $K_M$ values for ethanol can contribute to metabolism depending on the blood levels achieved after a particular dose of ethanol (2 g/kg in our work). Our results also pointed to the functional context for inclusion of these alcohol dehydrogenases in a module which, in the rat, under normal conditions rarely, if at all, experiences alcohol concentrations of 40 mM or higher. The same can be said for most humans and this module's involvement in generation and utilization of retinoic acid is another relevant component of our results.

We would suggest that the protocol we implemented that included QTL or GWAS of physiologic, pathologic and behavioral traits in animals, including humans, can bring credence to anticipated results and introduce unexpected but plausible systems genetic explanations of complex traits. Furthermore, since the RI rats utilized in this study have been stored for eternal use and we have procured a rich database of their attributes for public use, studies addressing alternative research questions, e.g., the influence of genes and/or modules on different phenotypes or how they predispose response to various environmental factors, can easily be employed in this panel and perhaps extrapolated to humans for valuable insights, The evidence for the possible contribution of *Aldh1a1* to acetate AUC although it was not included in a candidate module, must, however, temper the absolute utility of the coexpression approach and indicates that careful inspection of all forms of gene expression data in relationship to a given phenotype is still necessary to reach optimum conclusions. Additionally, it should be noted that

our study design included only male rats, and gender differences in alcohol metabolism have been reported [470].

In summary, here we sought to provide evidence that the association between alcohol metabolizing genes and AUD may be at least partly due to acetate. These genes are directly involved with the production of acetate. Our work demonstrated that differential expression of *Adh1* and *Adh4* and not *ALDH* isoforms, influences acetate exposure in rats. *ADH1* and *ADH4* have some of the strongest associations with AUD in humans and possess noncoding SNPs that influence expression. In addition, the genetic etiology of complex traits such as AUD are likely driven by gene regulation processes such as expression. Beyond the genetic support for acetate as the metabolic and biological link between the genome and AUD, we also provided two possible modes of action for acetate. Namely, acetate from ethanol reaches the brain and causes 1) hyperacetylation and corresponding gene expression changes, and 2) an increase in brain acetate metabolism for energy needs during development of AUD, rather than glucose metabolism. Future studies on acetate derived from alcohol are needed to clearly establish its potential role in AUD that ideally combine or consider its molecular, genetic and genomic aspects.

**Aim 2: Develop a machine learning algorithm for identifying polyA sites in the expressed transcriptome that utilizes DNA sequence and short read bulk high-throughput RNA-Seq (Chapter III)**

In Chapter II, we utilized gene level expression estimates when identifying candidate coexpression networks associated with a complex trait. In other words, this analysis did not distinguish between isoforms derived from the same gene. While grouping expression estimates by gene may possess inherent advantages, e.g., it likely yields more robust expression estimates,

is easily interpretable, many complex traits are understood to be modulated by only one isoform derived from a gene even when multiple isoforms of that gene are expressed. One post-transcriptional modification that combines aspects of expression and different transcript structures is alternative polyadenylation (APA). APA transcripts derived from the same gene, by definition, have different transcript structures. However, APA transcripts often share the same protein coding sequence and only differ in 3' UTR length. Regardless, APA transcripts can have distinct biological consequences; for example, expression of one APA isoform at the expense of another can reduce the overall expression level of the gene through differential miRNA binding between these isoforms.

APA is a pervasive gene regulation mechanism that may play a role alcohol related complex traits; however, current methods for identifying polyA sites and characterizing APA suffer several shortcomings. As a result, in Chapter III, we developed an algorithm to improve annotation of polyA sites in the expressed transcriptome whose results can be incorporated into downstream analyses such as that utilized in Chapter II.

The algorithm we developed, aptardi, utilized multiple omics sources and supervised machine learning. Specifically, we harnessed the information afforded by both DNA sequence and high-throughput RNA-Seq to predict polyA sites. For machine learning, we utilized a supervised paradigm where the labels were derived from the highly accurate PolyA-Seq data. The machine learning algorithm was a biLSTM, which takes into account the sequential nature of bins across a transcript when making predictions.

We first demonstrated that aptardi is broadly applicable. We did this by showing that a model trained on one dataset performs comparably on unseen data from difference sources. We then showed that aptardi outperforms current methods for identifying polyA sites. In particular,

we showed that applying aptardi to current transcriptome reconstruction pipelines improves polyA site annotation. We further showed that aptardi outperforms RNA-Seq based and DNA sequence-based algorithms for identifying polyA sites. Finally, we applied aptardi to a dataset where APA events were experimentally confirmed and showed that aptardi could successfully identify these APA sites in a sample specific manner (i.e., identified the sites in the knockdown but not the control) although they were not identified using current reconstruction methods.

We made aptardi publicly available as free, open-source software. Therefore, future research can use aptardi to study the impact of APA on various processes and diseases.

One limitation of aptardi is that it cannot distinguish transcript/polyA site pairs when multiple transcripts overlap a genomic region where it identified a polyA site. As a result, it enumerates all possible transcripts for a given polyA site when multiple transcripts are present. Future research to assign polyA sites to specific transcripts would enhance the algorithm. Another perceived limitation may be that aptardi makes predictions on 100 base bins (for positive predictions, it annotates the transcript stop site as the 3' most base in the given 100 base bin). However, this concern is partially mitigated because the precise location of polyA sites can "wiggle" by up to 30 nucleotides for what is considered a single isoform, i.e. not an APA event [178], and as such researchers often group polyA sites within 30 bases into a single site [208]. Furthermore, few 100 base bins contained multiple polyA sites (Supplementary Table 4). Finally, the manually engineered DNA sequence features may not apply to taxa outside of mammals [178] and will require further research for extension to other taxa, such as plans and insects.

**Aim 3: Characterize the alternative splicing and alternative polyadenylation transcriptional landscape in brain of the HXB/BXH recombinant inbred rat panel and assess its impact on predisposition to voluntary alcohol consumption (Chapter IV)**

Alternative splicing and APA greatly increase the transcriptomic diversity in eukaryotes. However, these events are often under-annotated in reference annotation, especially in model organisms such as rat. In addition, alternative splicing and alternative polyadenylation can vary based on physiological conditions, cell type, developmental, stage, sex, immune and disease state, immune response, inflammation, and viral infection. Moreover, brain exhibits the greatest diversity from these phenomena of all tissues. Finally, as previously delineated, previous research has implicated alternative splicing and APA in the genetic underpinnings of complex diseases such as alcohol related phenotypes.

In Chapter IV, we had two main goals 1) characterize the alternative splicing and APA transcriptional landscape in brain of the HXB/BXH RI rat panel and 2) identify candidate transcripts and transcript networks associated with voluntary alcohol consumption.

To generate a brain specific transcriptome of the HXB/BXH RI panel assessed for alternative splicing and APA, we applied StringTie (for alternative splicing) and aptardi (for APA) – the algorithm we developed Chapter III. We first established a filtering pipeline for identifying high quality transcripts when applying these algorithms. Of note, since this was the first time applying aptardi to data of this magnitude, and since aptardi cannot distinguish transcript/polyA site pairs for overlapping transcripts, we assessed how to filter potential false positive transcripts identified by aptardi. Others can utilize this pipeline when applying aptardi in their studies. This filtering was used to establish an expressed transcriptome. The number of isoforms was greatly increased in this transcriptome compared to the reference transcriptome,

supporting the notion that most genes undergo alternative splicing and APA and that these events are under-annotated. Supporting their genetic component, transcripts identified by StringTie and aptardi showed comparable heritability to those in reference annotation.

We then applied the statistical pipeline developed in Chapter II to identify candidate transcript coexpression networks. By doing this at the transcript level, we were able to resolve isoforms of genes, unlike the method employed previously. We successfully recapitulated results from previous studies examining genes associated with this phenotype while further advancing our knowledge by identifying specific isoforms of these genes responsible for the association. We likewise identified individual candidate transcripts associated with voluntary alcohol consumption, some of which included novel alternative splicing and APA transcripts identified through our computational approach.

## Future directions

### Aim 1: Develop a systems genetics pipeline for identifying networks of genes associated with a complex trait (Chapter II).

Our use-all-data, systems genetics approach to identify networks of genes associated with a complex trait was successfully validated by identifying a candidate module containing the *ADH* genes, which are well-known to influence the alcohol metabolism phenotypes. As a result, we suggest researchers can apply our methodology to other phenotypes to likewise investigate networks of genes associated with a complex trait. Another interesting aspect of this work was inclusion of acetate as a phenotype. The same *ADH* containing candidate module was associated with this phenotype. Much of the work focused on explaining the molecular pathways influenced by the alcohol metabolizing genes that lead to their association with AUD have focused on acetaldehyde. Yet results have been inconsistent, and we provided a possible explanation for

how acetate may be a mediator of AUD. Follow up experiments on the role of acetate in AUD, with a particular focus on a design that allows for differentiating the contributions of acetate and acetaldehyde, would likely lead to better insights into a possible role for acetate.

**Aim 2: Develop a machine learning algorithm for identifying polyA sites in the expressed transcriptome that utilizes DNA sequence and short read bulk high-throughput RNA-Seq (Chapter III)**

Two divergent future directions could be pursued in regard to Aim 2: 1) improvement of the algorithm and 2) use of the algorithm to gain biological insight. To improve the algorithm, several avenues are available. For instance, aptardi cannot distinguish transcript/polyA site pairs for overlapping transcripts. The ability to assign a single polyA site to a single transcript would therefore be beneficial but may require alternative datasets. Another possibility is to cultivate other features and potentially perform automated feature engineering using more sophisticated machine learning methods. Likewise, alternative hyperparameters and more complex machine learning models could be explored. Reducing the bin size that aptardi identifies polyA sites within would increase the resolution of polyA sites. Smoothing of the RNA-Seq data to reduce false positives resulting from uneven coverage changes due to sequencing phenomena such as GC biases could also improve performance. Future research could also use the algorithm to better characterize the expressed transcriptome by improving identification of 3' ends of transcripts. In Chapter IV, we demonstrated how aptardi could be incorporated into transcriptome assembly. By improving the resolution transcript structures of genes using aptardi, better expression estimates of specific isoforms can be obtained and utilized to ascertain the role of specific isoform expression on the phenotype. Likewise, improved annotation of the expressed transcriptome may provide greater detail into the exact transcript isoforms responsible for a

phenotype. This could present new avenues for therapeutic targets, e.g., the discovery of an aptardi transcript with a longer 3' UTR than previously annotated could enable therapies targeted at newly acquired miRNA binding sites. Furthermore, assembling sample specific transcriptomes with aptardi may uncover differences in transcripts that could serve as a biomarker of a given disease state, e.g., shortening or lengthening of transcript 3' UTRs as identified by aptardi could indicate cancer vs healthy individuals. Alternatively, understanding differences in transcript structures from aptardi analysis or expression levels may provide insight into individual differences in response to a therapy.

**Aim 3: Characterize the alternative splicing and alternative polyadenylation transcriptional landscape in brain of the HXB/BXH recombinant inbred rat panel and assess its impact on predisposition to voluntary alcohol consumption (Chapter IV)**

Future research should focus on several key areas. One area would be to experimentally validate the computationally derived transcript structures identified here that were in the candidate transcript coexpression module (e.g., *Ift81* and *Mapkapk5*) or individual candidate transcripts (e.g., *Map3k7* and *Aldh1a7*). For those with different 3' ends compared to existing reference transcripts as identified by aptardi (e.g., *Ift81* and *Aldh1a7*), the gain and/or loss of potential microRNA binding sites may be of interest and the biological consequences elucidated with respect to voluntary alcohol consumption. For those with different exon structures (i.e., *Mapkapk5* and *Map3k7*), the impact on the encoded protein may provide insight into their association with voluntary alcohol consumption.

Other work could address the robustness of the candidate individual transcripts using functional validation studies and/or increasing sample size. Likewise, ascertaining the relatedness of the module components and casual relationships to each other and to voluntary

alcohol consumption would likely yield important insights. Another potentially interesting study would be to evaluate these candidate modules and candidate individual transcripts after exposure to alcohol (rather than as predisposing factors).

## Concluding remarks

The genetic architecture of complex traits is, as the name suggests, complex. In this work we sought to better understand the connection between genotype and phenotype by including the information afforded by RNA expression data. Specifically, we first validated a methodology for identifying networks of coexpressed genes associated with a complex trait. Networks of genes provide information on the biological processes underlying a phenotype. Next, we developed a machine learning algorithm for identifying the 3' ends, or polyA sites, of transcripts in the expressed transcriptome. Post transcriptional modification of transcripts allows for dynamic usage of 3' ends, thereby necessitating sample specific analysis to uncover the impact of this on phenotypes. Finally, we applied these methods to gain additional insight into the genetic underpinnings of voluntary alcohol consumption.

## REFERENCES


1. Organization, W.H., *Global status report on alcohol and health 2018.* 2018.

2. Sacks, J.J., et al., *2010 National and State Costs of Excessive Alcohol Consumption.* Am J Prev Med, 2015. **49**(5): p. e73-e79.

3. Association, A.P., *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition.* American Psychiatric Association, 2013.

4. Grant, B.F., et al., *Epidemiology of DSM-5 Alcohol Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions III.* JAMA Psychiatry, 2015. **72**(8): p. 757-766.

5. Bynum, W.F., *Alcoholism and degeneration in 19th century European medicine and psychiatry.* British journal of addiction, 1984. **79**(1): p. 59-70.

6. Cotton, N.S., *The familial incidence of alcoholism: a review.* Journal of Studies on Alcohol, 1979. **40**(1): p. 89-116.

7. Kaij, L. and D. Rosenthal, *Alcoholism in Twins. Studies on the Etiology and Sequels of Abuse of Alcohol.* The Journal of Nervous and Mental Disease, 1961. **133**(3): p. 272.

8. McGue, M., R.W. Pickens, and D.S. Svikis, *Sex and age effects on the inheritance of alcohol problems: a twin study.* Journal of abnormal psychology, 1992. **101**(1): p. 3-17.

9. Reed, T., et al., *Genetic predisposition to organ-specific endpoints of alcoholism.* Alcoholism: Clinical and Experimental Research, 1996. **20**(9): p. 1528-1533.

10. Heath, A.C., K.K. Bucholz, and P. Madden, *Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men.* Psychological Medicine, 1997. **27**(6): p. 1381-1396.

11. Prescott, C.A. and K.S. Kendler, *Genetic and Environmental Contributions to Alcohol Abuse and Dependence in a Population-Based Sample of Male Twins.* American Journal of Psychiatry, 1999. **156**(1): p. 34-40.

12. Knopik, V.S., et al., *Genetic effects on alcohol dependence risk: re-evaluating the importance of psychiatric and other heritable risk factors.* Psychological Medicine, 2004. **34**(8): p. 1519-1530.

13. Magnusson, Å., et al., *Familial influence and childhood trauma in female alcoholism.* Psychological Medicine, 2012. **42**(2): p. 381-389.

14. Goodwin, D.W., et al., *Alcohol problems in adoptees raised apart from alcoholic biological parents.* Archives of General Psychiatry, 1973. **28**(2): p. 238-243.

15. Bohman, M., S. Sigvardsson, and C.R. Cloninger, *Maternal inheritance of alcohol abuse. Cross-fostering analysis of adopted women.* Archives of General Psychiatry, 1981. **38**(9): p. 965-969.

16. Cloninger, C.R., M. Bohman, and S. Sigvardsson, *Inheritance of alcohol abuse. Cross-fostering analysis of adopted men.* Archives of General Psychiatry, 1981. **38**(8): p. 861-868.

17. Cadoret, R.J., E. Troughton, and T.W. O'Gorman, *Genetic and environmental factors in alcohol abuse and antisocial personality.* Journal of Studies on Alcohol, 1987. **48**(1): p. 1-8.

18. Sigvardsson, S., M. Bohman, and C.R. Cloninger, *Replication of the Stockholm Adoption Study of alcoholism. Confirmatory cross-fostering analysis.* Archives of General Psychiatry, 1996. **53**(8): p. 681-687.

19. Verhulst, B., M.C. Neale, and K.S. Kendler, *The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies.* Psychological Medicine, 2015. **45**(5): p. 1061-1072.

20. World, A.H.A.H.R. and 1995, *Genetic influences on alcoholism risk.* pubs.niaaa.nih.gov.

21. Prescott, C.A., et al., *Gender and genetic vulnerability to alcoholism*.

22. Dick, D.M. and L.J. Bierut, *The genetics of alcohol dependence.* Current psychiatry reports, 2006. **8**(2): p. 151-157.

23. Edenberg, H.J. and T. Foroud, *Genetics and alcoholism.* Nature Reviews Gastroenterology & Hepatology, 2013. **10**(8): p. 487.

24. Kimura, M. and S. Higuchi, *Genetics of alcohol dependence.* Psychiatry and clinical neurosciences, 2011. **65**(3): p. 213-225.

25. Edenberg, H.J. and T. Foroud, *Review: The genetics of alcoholism: identifying specific genes through family studies.* Addiction Biology, 2006. **11**(3-4): p. 386.

26. Enoch, M.A. and D. Goldman, *The genetics of alcoholism and alcohol abuse.* 2001. **3**(2): p. 144.

27. Hines, L.M., L. Ray, and K. Hutchison, *Alcoholism: the dissection for endophenotypes.* Psychophysiology, 2005. **51**(12): p. 1337.

28. Parker, C.C., R. Lusk, and L.M. Saba, *Alcohol Sensitivity as an Endophenotype of Alcohol Use Disorder: Exploring Its Translational Utility between Rodents and Humans.* Brain Sci, 2020. **10**(10).

29. Hart, A.B. and H.R. Kranzler, *Alcohol Dependence Genetics: Lessons Learned From Genome-Wide Association Studies (GWAS) and Post-GWAS Analyses.* Alcoholism: Clinical and Experimental Research, 2015. **39**(8): p. 1312-1327.

30. Olfson, E. and L.J. Bierut, *Convergence of genome-wide association and candidate gene studies for alcoholism.* Alcoholism: Clinical and Experimental Research, 2012. **36**(12): p. 2086-2094.

31. Deak, J.D., A.P. Miller, and I.R. Gizer, *Genetics of alcohol use disorder: a review.* Current opinion in psychology, 2019. **27**: p. 56-61.

32. Edenberg, H.J. and J.N. McClintick, *Alcohol Dehydrogenases, Aldehyde Dehydrogenases, and Alcohol Use Disorders: A Critical Review.* Alcoholism: Clinical and Experimental Research, 2018. **42**(12): p. 2281-2297.

33. Thomasson, H.R., et al., *Alcohol and aldehyde dehydrogenase genotypes and alcoholism in Chinese men.* American journal of human genetics, 1991. **48**(4): p. 677-681.

34. Bierut, L.J., et al., *ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry.* Molecular psychiatry, 2012. **17**(4): p. 445-450.

35. Gelernter, J., et al., *Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci.* Molecular psychiatry, 2014. **19**(1): p. 41-49.

36. Gelernter, J., et al., *Genomewide Association Study of Alcohol Dependence and Related Traits in a Thai Population.* Alcoholism: Clinical and Experimental Research, 2018. **42**(5): p. 861-868.

37. Park, B.L., et al., *Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication.* Human Genetics, 2013. **132**(6): p. 657-668.

38. Zintzaras, E., et al., *Do alcohol-metabolizing enzyme gene polymorphisms increase the risk of alcoholism and alcoholic liver disease?* Hepatology, 2006. **43**(2): p. 352-361.

39. Walters, R.K., et al., *Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders.* Nature Neuroscience, 2018. **21**(12): p. 1656-1669.

40. Sanchez-Roige, S., et al., *Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts.* American Journal of Psychiatry, 2018. **176**(2): p. 107-118.

41. Frank, J., et al., *Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster.* Addiction Biology, 2012. **17**(1): p. 171-180.

42. Ramchandani, V.A., W.F. Bosron, and T.K. Li, *Research advances in ethanol metabolism.* Pathologie Biologie, 2001. **49**(9): p. 676.

43. Norberg, A., et al., *Role of variability in explaining ethanol pharmacokinetics: research and forensic applications.* Clinical Pharmacokinetics, 2003. **42**(1): p. 1-31.

44. Zakhari, S., *Overview: how is alcohol metabolized by the body?* Alcohol Research & Health, 2006.

45. Guindalini, C., et al., *Association of genetic variants in alcohol dehydrogenase 4 with alcohol dependence in Brazilian patients.* American Journal of Psychiatry, 2005. **162**(5): p. 1005-1007.

46. Cederbaum, A.I., *Alcohol metabolism.* Clinics in liver disease, 2012. **16**(4): p. 667-685.

47. Bosron, W.F., T. Ehrig, and T.K. Li, *Genetic factors in alcohol metabolism and alcoholism.* Alcohol Metabolism, Alcohol Intolerance, and Alcoholism, 1993: p. 107.

48. Edenberg, H.J. and T. Foroud, *Genetics of alcoholism.* Handbook of Clinical Neurology, 2014. **125**: p. 561-571.

49. Yin, S.-J., et al., *Human Alcohol Dehydrogenase Family*. 1999, Springer, Boston, MA: Boston, MA. p. 265-274.

50. Edenberg, H.J., *The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants.* Alcohol and Aldehyde Metabolizing Systems, 2007: p. 335.

51. Chen, C.C., et al., *Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism.* American journal of human genetics, 1999. **65**(3): p. 795-807.

52. Bosron, W.F. and T.K. Li, *Genetic polymorphism of human liver alcohol and aldehyde dehydrogenases, and their relationship to alcohol metabolism and alcoholism.* Hepatology, 1986. **6**(3): p. 502-510.

53. Hurley, T.D. and H.J. Edenberg, *Genes encoding enzymes involved in ethanol metabolism.* Alcohol research: current reviews, 2012. **34**(3): p. 339-344.

54. Edenberg, H. and W.F. Bosron, *Alcohol Dehydrogenases*. 2017, Elsevier Inc. p. 126-145.

55. Edenberg, H.J., et al., *Association of alcohol dehydrogenase genes with alcohol dependence: a comprehensive analysis.* Human Molecular Genetics, 2006. **15**(9): p. 1539-1549.

56. Jackson, B., et al., *Update on the aldehyde dehydrogenase gene ( ALDH ) superfamily.* Human Genomics, 2011. **5**(4): p. 1-21.

57. Vasiliou, V., A.P. Pharmacology, and 2000, *Polymorphisms of human aldehyde dehydrogenases.* karger.com.

58. Vasiliou, D.V., A. Pappa, and T. Estey, *Role of Human Aldehyde Dehydrogenases in Endobiotic and Xenobiotic Metabolism.* Drug Metabolism Reviews, 2004. **36**(2): p. 279-299.

59. Vasiliou, V., et al., *Aldehyde dehydrogenases: from eye crystallins to metabolic disease and cancer stem cells.* Elsevier.

60. Stagos, D., et al., *Aldehyde Dehydrogenase 1B1: Molecular Cloning and Characterization of a Novel Mitochondrial Acetaldehyde-Metabolizing Enzyme.* Drug Metabolism and Disposition, 2010. **38**(10): p. 1679-1687.

61. Klyosov, A.A., *Kinetics and Specificity of Human Liver Aldehyde Dehydrogenases toward Aliphatic, Aromatic, and Fused Polycyclic Aldehydes†.* Biochemistry, 1996. **35**(14): p. 4457-4467.

62. Singh, S., et al., *ALDH1B1 links alcohol consumption and diabetes.* Biochemical and Biophysical Research Communications, 2015. **463**(4): p. 768-773.

63. Reich, T., et al., *Genome-wide search for genes affecting the risk for alcohol dependence.* American journal of medical genetics, 1998. **81**(3): p. 207-215.

64. Long, J.C., et al., *Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population.* American journal of medical genetics, 1998. **81**(3): p. 216-221.

65. Prescott, C.A., et al., *Genomewide linkage study in the Irish affected sib pair study of alcohol dependence: evidence for a susceptibility region for symptoms of alcohol dependence on chromosome 4.* Molecular psychiatry, 2006. **11**(6): p. 603-611.

66. Williams, J.T., et al., *Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. II. Alcoholism and event-related potentials.* American journal of human genetics, 1999. **65**(4): p. 1148-1160.

67. Saccone, N.L., et al., *A genome screen of maximum number of drinks as an alcoholism phenotype.* American journal of medical genetics, 2000. **96**(5): p. 632-637.

68. Zinn-Justin, A. and L. Abel, *Genome search for alcohol dependence using the weighted pairwise correlation linkage method: interesting findings on chromosome 4.* Genetic epidemiology, 1999. **17 Suppl 1**(S1): p. S421-6.

69. Ehlers, C.L., et al., *Association of ALDH1 promoter polymorphisms with alcohol-related phenotypes in southwest California Indians.* Alcoholism: Clinical and Experimental Research, 2004. **28**(10): p. 1481-1486.

70. Hill, S.Y., et al., *A genome wide search for alcoholism susceptibility genes.* American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 2004. **128B**(1): p. 102-113.

71. Luo, X., et al., *ADH4 gene variation is associated with alcohol and drug dependence: results from family controlled and population-structured association studies.* Pharmacogenetics and Genomics, 2005. **15**(11): p. 755-768.

72. Kuo, P.-H., et al., *Association of ADH and ALDH genes with alcohol dependence in the Irish Affected Sib Pair Study of alcohol dependence (IASPSAD) sample.* Alcoholism: Clinical and Experimental Research, 2008. **32**(5): p. 785-795.

73. Li, D., H. Zhao, and J. Gelernter, *Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases.* Biological Psychiatry, 2011. **70**(6): p. 504-512.

74. Li, D., H. Zhao, and J. Gelernter, *Strong protective effect of the aldehyde dehydrogenase gene (ALDH2) 504lys (*2) allele against alcoholism and alcohol-induced medical diseases in Asians.* Human Genetics, 2012. **131**(5): p. 725-737.

75. Li, D., H. Zhao, and J. Gelernter, *Further clarification of the contribution of the ADH1C gene to vulnerability of alcoholism and selected liver diseases.* Human Genetics, 2012. **131**(8): p. 1361-1374.

76. Quillen, E.E., et al., *ALDH2 is associated to alcohol dependence and is the major genetic determinant of "daily maximum drinks" in a GWAS study of an isolated rural Chinese sample.* American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 2014. **165B**(2): p. 103-110.

77. Spence, J.P., et al., *Evaluation of aldehyde dehydrogenase 1 promoter polymorphisms identified in human populations.* Alcoholism: Clinical and Experimental Research, 2003. **27**(9): p. 1389-1394.

78. Sherva, R., et al., *Associations and interactions between SNPs in the alcohol metabolizing genes and alcoholism phenotypes in European Americans.* Alcoholism: Clinical and Experimental Research, 2009. **33**(5): p. 848-857.

79. Agrawal, A., et al., *A candidate gene association study of alcohol consumption in young women.* Alcoholism: Clinical and Experimental Research, 2011. **35**(3): p. 550-558.

80. Husemoen, L.L.N., et al., *The association of ADH and ALDH gene variants with alcohol drinking habits and cardiovascular disease risk factors.* Alcoholism: Clinical and Experimental Research, 2008. **32**(11): p. 1984-1991.

81. Linneberg, A., et al., *Genetic determinants of both ethanol and acetaldehyde metabolism influence alcohol hypersensitivity and drinking behaviour among Scandinavians.* Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology, 2010. **40**(1): p. 123-130.

82. Bjerregaard, P., et al., *Genetic variation in alcohol metabolizing enzymes among Inuit and its relation to drinking patterns.* Drug and alcohol dependence, 2014. **144**: p. 239-244.

83. Way, M.J., et al., *Genetic variants in ALDH1B1 and alcohol dependence risk in a British and Irish population: A bioinformatic and genetic study.* PLoS ONE, 2017. **12**(6): p. e0177009.

84. Chi, Y.C., et al., *Modeling of Human Hepatic and Gastrointestinal Ethanol Metabolism with Kinetic-Mechanism-Based Full-Rate Equations of the Component Alcohol Dehydrogenase Isozymes and Allozymes.* Chem Res Toxicol, 2018. **31**(7): p. 556-569.

85. Osier, M.V., et al., *A proline-threonine substitution in codon 351 of ADH1C is common in Native Americans.* Alcoholism: Clinical and Experimental Research, 2002. **26**(12): p. 1759-1763.

86. Strömberg, P., et al., *Identification and characterisation of two allelic forms of human alcohol dehydrogenase 2.* Cellular and molecular life sciences : CMLS, 2002. **59**(3): p. 552-559.

87. Chou, W.Y., et al., *An A/G polymorphism in the promoter of mitochondrial aldehyde dehydrogenase (ALDH2): effects of the sequence variant on transcription factor binding and promoter strength.* Alcoholism: Clinical and Experimental Research, 1999. **23**(6): p. 963-968.

88. Crabb, D.W., et al., *Genotypes for aldehyde dehydrogenase deficiency and alcohol sensitivity. The inactive ALDH2(2) allele is dominant.* Journal of Clinical Investigation, 1989. **83**(1): p. 314-316.

89. Zhou, J. and H. Weiner, *Basis for half-of-the-site reactivity and the dominance of the K487 oriental subunit over the E487 subunit in heterotetrameric human liver mitochondrial aldehyde dehydrogenase.* Biochemistry, 2000. **39**(39): p. 12019-12024.

90. Xiao, Q., H. Weiner, and D.W. Crabb, *The mutation in the mitochondrial aldehyde dehydrogenase (ALDH2) gene responsible for alcohol-induced flushing increases turnover of the enzyme tetramers in a dominant fashion.* Journal of Clinical Investigation, 1996. **98**(9): p. 2027-2032.

91. Hsu, L.C. and W.C. Chang, *Cloning and characterization of a new functional human aldehyde dehydrogenase gene.* Journal of Biological Chemistry, 1991. **266**(19): p. 12257-12265.

92. Sherman, D., et al., *Diverse polymorphism within a short coding region of the human aldehyde dehydrogenase-5 (ALDH5) gene.* Human Genetics, 1993. **92**(5): p. 477-480.

93. Chen, H.-J., H. Tian, and H.J. Edenberg, *Natural haplotypes in the regulatory sequences affect human alcohol dehydrogenase 1C (ADH1C) gene expression.* Human mutation, 2005. **25**(2): p. 150-155.

94. Pochareddy, S. and H.J. Edenberg, *Variation in the ADH1B proximal promoter affects expression.* Chemico-biological interactions, 2011. **191**(1-3): p. 38-41.

95. Edenberg, H.J., R.E. Jerome, and M. Li, *Polymorphism of the human alcohol dehydrogenase 4 (ADH4) promoter affects gene expression.* Pharmacogenetics, 1999. **9**(1): p. 25-30.

96. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Research, 2013. **42**(D1): p. D1001-D1006.

97. Pickrell, J.K., *Joint analysis of functional genomic data and genome-wide association studies of 18 human traits.* American journal of human genetics, 2014. **94**(4): p. 559-573.

98. Li, Y.I., et al., *RNA splicing is a primary link between genetic variation and disease.* Science, 2016. **352**(6285): p. 600-604.

99. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.

100. Luo, X., et al., *ADH4 gene variation is associated with alcohol dependence and drug dependence in European Americans: results from HWD tests and case-control association studies.* Neuropsychopharmacology, 2006. **31**(5): p. 1085-95.

101. Edenberg, H.J., et al., *Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations.* Am J Hum Genet, 2004. **74**(4): p. 705-14.

102. Tian, B. and J.L. Manley, *Alternative polyadenylation of mRNA precursors.* Nature reviews. Molecular cell biology, 2017. **18**(1): p. 18-30.

103. Yong, H.-S.Y.a.J., *Alternative Polyadenylation of mRNAs: 3′-Untranslated Region Matters in Gene Expression.* Molecules and cells, 2016. **39**(4): p. 281-285.

104. Di Giammartino, D.C., K. Nishida, and J.L. Manley, *Mechanisms and consequences of alternative polyadenylation.* Molecular cell, 2011. **43**(6): p. 853-866.

105. Chen, H.J., H. Tian, and H.J. Edenberg, *Natural haplotypes in the regulatory sequences affect human alcohol dehydrogenase 1C (ADH1C) gene expression.* Hum Mutat, 2005. **25**(2): p. 150-5.

106. Pochareddy, S. and H.J. Edenberg, *Identification of a FOXA-dependent enhancer of human alcohol dehydrogenase 4 (ADH4).* Gene, 2010. **460**(1-2): p. 1-7.

107. Grant, J.D., et al., *Alcohol Consumption Indices of Genetic Risk for Alcohol Dependence.* Biological Psychiatry, 2009. **66**(8): p. 795-800.

108. Kendler, K.S., et al., *A Population-Based Twin Study of Alcoholism in Women.* JAMA, 1992. **268**(14): p. 1877.

109. Collaborators, T.U.B.o.D., *The State of US Health, 1990-2016: Burden of Diseases, Injuries, and Risk Factors Among US States.* JAMA, 2018. **319**(14): p. 1444-1472.

110. Kapoor, M., et al., *A meta-analysis of two genome-wide association studies to identify novel loci for maximum number of alcoholic drinks.* Human Genetics, 2013. **132**(10): p. 1141-1151.

111. Murphy, J.M., et al., *Phenotypic and genotypic characterization of the Indiana University rat lines selectively bred for high and low alcohol preference.* Behavior genetics, 2002. **32**(5): p. 363-388.

112. Swan, G.E., et al., *Smoking and alcohol consumption in adult male twins: Genetic heritability and shared environmental influences.* Journal of Substance Abuse, 1990. **2**(1): p. 39-50.

113. Ehlers, C.L., et al., *A comparison of selected quantitative trait loci associated with alcohol use phenotypes in humans and mouse models.* Addiction Biology, 2010. **15**(2): p. 185-199.

114. Hirschhorn, J.N., et al., *A comprehensive review of genetic association studies.* Genetics in Medicine, 2002. **4**(2): p. 45-61.

115. Palmer, R.H.C., et al., *The genetics of alcohol dependence: advancing towards systems-based approaches.* Drug and alcohol dependence, 2012. **125**(3): p. 179-191.

116. *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

117. Litten, R.Z., et al., *Medications development to treat alcohol dependence: a vision for the next decade.* Addict Biol, 2012. **17**(3): p. 513-27.

118. Swift, R.M. and E.R. Aston, *Pharmacotherapy for alcohol use disorder: current and emerging therapies.* Harv Rev Psychiatry, 2015. **23**(2): p. 122-33.

119. Volpicelli, J.R., et al., *Naltrexone in the treatment of alcohol dependence.* Arch Gen Psychiatry, 1992. **49**(11): p. 876-80.

120. Harris, B.R., et al., *Acamprosate inhibits the binding and neurotoxic effects of trans-ACPD, suggesting a novel site of action at metabotropic glutamate receptors.* Alcohol Clin Exp Res, 2002. **26**(12): p. 1779-93.

121. Foroud, T. and H.J. Edenberg, *Genetic research.* … Abuse and Alcoholism, 2010.

122. Yin, S.J., *Alcohol dehydrogenase: enzymology and metabolism.* Alcohol and alcoholism (Oxford, Oxfordshire). Supplement, 1993. **2**: p. 113-119.

123. Salvatore, J.E., et al., *Beyond genome-wide significance: integrative approaches to the interpretation and extension of GWAS findings for alcohol use disorder.* Addiction Biology, 2019. **24**(2): p. 275-289.

124. McBride, W.J. and T.K. Li, *Animal models of alcoholism: neurobiology of high alcohol-drinking behavior in rodents.* Critical reviews in neurobiology, 1998. **12**(4): p. 339-369.

125. Foroud, T., H.J. Edenberg, and J.C. Crabbe, *Genetic research: who is at risk for alcoholism.* Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, 2010. **33**(1-2): p. 64-75.

126. Silver, L.M., *Mouse genetics : concepts and applications*. 1995, New York: Oxford University Press. xiii, 362 p.

127. Toth, L.A., R.A. Trammell, and R.W. Williams, *Mapping complex traits using families of recombinant inbred strains: an overview and example of mapping susceptibility to Candida albicans induced illness phenotypes.* Pathogens and Disease, 2014. **71**(2): p. 234-248.

128. Belknap, J.K., *Effect of Within-Strain Sample Size on QTL Detection and Mapping Using Recombinant Inbred Mouse Strains.* Behavior Genetics, 1998. **28**(1): p. 29-38.

129. Tabakoff, B., et al., *Genetical genomic determinants of alcohol consumption in rats and humans.* BMC biology, 2009. **7**(1): p. 70.

130. Saba, L.M., et al., *A systems genetic analysis of alcohol drinking by mice, rats and men: influence of brain GABAergic transmission.* Neuropharmacology, 2011. **60**(7-8): p. 1269-1280.

131. Vanderlinden, L.A., et al., *Is the alcohol deprivation effect genetically mediated? Studies with HXB/BXH recombinant inbred rat strains.* Alcohol Clin Exp Res, 2014. **38**(7): p. 2148-57.

132. Saba, L.M., et al., *The sequenced rat brain transcriptome – its use in identifying networks predisposing alcohol consumption.* The FEBS Journal, 2015. **282**(18): p. 3556-3578.

133. Hoffman, P.L., et al., *Voluntary exposure to a toxin: the genetic influence on ethanol consumption.* Mamm Genome, 2018. **29**(1-2): p. 128-140.

134. Printz, M.P., et al., *Genetic Models in Applied Physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics.* 2003. **94**(6): p. 2510-2522.

135. Simonis, M., S.S. Atanur, and S. Linsen, *Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel.* 2012.

136. Blizard, D.A., *Recombinant-inbred strains: general methodological considerations relevant to the study of complex characters.* Behav Genet, 1992. **22**(6): p. 621-33.

137. Tabakoff, B., et al., *Genetical genomic determinants of alcohol consumption in rats and humans.* BMC Biol, 2009. **7**: p. 70.

138. Vanderlinden, L.A., et al., *Whole brain and brain regional coexpression network interactions associated with predisposition to alcohol consumption.* PLoS One, 2013. **8**(7): p. e68878.

139. Kunes, J., et al., *Use of recombinant inbred strains for evaluation of intermediate phenotypes in spontaneous hypertension.* Clin Exp Pharmacol Physiol, 1994. **21**(11): p. 903-6.

140. Bielavska, E., et al., *Genome scanning of the HXB/BXH sets of recombinant inbred strains of the rat for quantitative trait loci associated with conditioned taste aversion.* Behav Genet, 2002. **32**(1): p. 51-6.

141. Conti, L.H., et al., *Identification of quantitative trait Loci for anxiety and locomotion phenotypes in rat recombinant inbred strains.* Behav Genet, 2004. **34**(1): p. 93-103.

142. Pravenec, M., et al., *Genetic analysis of "metabolic syndrome" in the spontaneously hypertensive rat.* Physiol Res, 2004. **53 Suppl 1**: p. S15-22.

143. Jornvall, H. and J.O. Hoog, *Nomenclature of alcohol dehydrogenases.* Alcohol Alcohol, 1995. **30**(2): p. 153-61.

144. Duester, G., et al., *Recommended nomenclature for the vertebrate alcohol dehydrogenase gene family.* Biochemical pharmacology, 1999. **58**(3): p. 389-395.

145. Svensson, S., P. Strömberg, and J.O. Höög, *A Novel Subtype of Class II Alcohol Dehydrogenase in Rodents: UNIQUE PRO47 and SER182 MODULATES HYDRIDE TRANSFER IN THE MOUSE ENZYME.* Journal of Biological Chemistry, 1999. **274**(42): p. 29712-29719.

146. Höög, J.-O. and L.J. Ostberg, *Mammalian alcohol dehydrogenases – A comparative investigation at gene and protein levels.* Chemico-biological interactions, 2011. **191**(1-3): p. 2-7.

147. Stranger, B.E., E.A. Stahl, and T. Raj, *Progress and promise of genome-wide association studies for human complex trait genetics.* Genetics, 2011. **187**(2): p. 367-83.

148. Beutler, B., et al., *GENETIC ANALYSIS OF HOST RESISTANCE: Toll-Like Receptor Signaling and Immunity at Large.* Annual Review of Immunology, 2006. **24**(1): p. 353-389.

149. Beutler, B., *Immunology, phenotype first. Preface.* Curr Top Microbiol Immunol, 2008. **321**: p. v-viii.

150. Moresco, E.M., X. Li, and B. Beutler, *Going forward with genetics: recent technological advances and forward genetics in mice.* Am J Pathol, 2013. **182**(5): p. 1462-73.

151. van der Sijde, M.R., A. Ng, and J. Fu, *Systems genetics: From GWAS to disease pathways.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2014. **1842**(10): p. 1903-1909.

152. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* Nat Rev Genet, 2014. **15**(1): p. 34-48.

153. Du, Q., et al., *Genetic architecture of growth traits in Populus revealed by integrated quantitative trait locus (QTL) analysis and association studies.* New Phytol, 2016. **209**(3): p. 1067-82.

154. Arnone, M.I. and E.H. Davidson, *The hardwiring of development: organization and function of genomic regulatory systems.* Development, 1997. **124**(10): p. 1851-64.

155. D'Haeseleer, P., S. Liang, and R. Somogyi, *Genetic network inference: from co-expression clustering to reverse engineering.* Bioinformatics, 2000. **16**(8): p. 707-26.

156. Zhao, W., et al., *Weighted Gene Coexpression Network Analysis: State of the Art.* Journal of Biopharmaceutical Statistics, 2010. **20**(2): p. 281-300.

157. Fuller, T.F., et al., *Weighted gene coexpression network analysis strategies applied to mouse weight.* Mamm Genome, 2007. **18**(6-7): p. 463-72.

158. Zhang, B. and S. Horvath, *A General Framework for Weighted Gene Co-Expression Network Analysis.* Statistical Applications in Genetics and Molecular Biology, 2005. **4**(1).

159. Weiss, J.N., et al., *"Good Enough Solutions" and the Genetics of Complex Diseases.* Circulation Research, 2012. **111**(4): p. 493-504.

160. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.

161. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.

162. Yip, A.M. and S. Horvath, *Gene network interconnectedness and the generalized topological overlap measure.* BMC Bioinformatics, 2007. **8**: p. 22.

163. Shi, Y., *Alternative polyadenylation: new insights from global analyses.* RNA (New York, N.Y.), 2012. **18**(12): p. 2105-2117.

164. Yeh, H.-S. and J. Yong, *Alternative Polyadenylation of mRNAs: 3'-Untranslated Region Matters in Gene Expression.* Molecules and cells, 2016. **39**(4): p. 281-285.

165. Jan, C.H., et al., *Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs.* Nature, 2011. **469**(7328): p. 97-101.

166. Mangone, M., et al., *The landscape of C. elegans 3'UTRs.* Science, 2010. **329**(5990): p. 432-5.

167. Ozsolak, F., et al., *Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation.* Cell, 2010. **143**(6): p. 1018-1029.

168. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA, 2011. **17**(4): p. 761-72.

169. Yoon, O.K., et al., *Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells.* PLoS Genetics, 2012. **8**(8): p. e1002882.

170. Manning, K.S. and T.A. Cooper, *The roles of RNA processing in translating genotype to phenotype.* Nature reviews. Molecular cell biology, 2016. **18**(2): p. 102-114.

171. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature Biotechnology, 2010. **28**(5): p. 511-515.

172. Fabian, M.R., N. Sonenberg, and W. Filipowicz, *Regulation of mRNA Translation and Stability by microRNAs.* Annual Review of Biochemistry, 2010. **79**(1): p. 351-379.

173. Sandberg, R., et al., *Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites.* Science, 2008. **320**(5883): p. 1643-7.

174. Ji, Z., et al., *Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development.* Proc Natl Acad Sci U S A, 2009. **106**(17): p. 7028-33.

175. Miranda, R.C., et al., *MicroRNAs: master regulators of ethanol abuse and toxicity?* Alcoholism: Clinical and Experimental Research, 2010. **34**(4): p. 575-587.

176. Pietrzykowski, A.Z., et al., *Posttranscriptional Regulation of BK Channel Splice Variant Stability by miR-9 Underlies Neuroadaptation to Alcohol.* Neuron, 2008. **59**(2): p. 274-287.

177. Civelek, M. and A.J. Lusis, *Systems genetics approaches to understand complex traits.* nature.com, 2014.

178. Tian, B. and J.H. Graber, *Signals for pre-mRNA cleavage and polyadenylation.* Wiley Interdisciplinary Reviews: RNA, 2012. **3**(3): p. 385-396.

179. Arefeen, A., X. Xiao, and T. Jiang, *DeepPASTA: deep neural network based polyadenylation site analysis.* Bioinformatics, 2019. **35**(22): p. 4577-4585.

180. Magana-Mora, A., M. Kalkatawi, and V.B. Bajic, *Omni-PolyA: a method and tool for accurate recognition of Poly(A) signals in human genomic DNA.* BMC Genomics, 2017. **18**(1): p. 620.

181. Leung, M.K.K., A. Delong, and B.J. Frey, *Inference of the human polyadenylation code.* Bioinformatics, 2018. **34**(17): p. 2889-2898.

182. Millevoi, S. and S. Vagner, *Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation.* Nucleic Acids Res, 2010. **38**(9): p. 2757-74.

183. Sanfilippo, P., J. Wen, and E.C. Lai, *Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species.* Genome biology, 2017. **18**(1): p. 229.

184. Zhang, H., J.Y. Lee, and B. Tian, *Biased alternative polyadenylation in human tissues.* Genome Biol, 2005. **6**(12): p. R100.

185. Beaudoing, E. and D. Gautheret, *Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data.* Genome Res, 2001. **11**(9): p. 1520-6.

186. Lenhard, B., A. Sandelin, and P. Carninci, *Metazoan promoters: emerging characteristics and insights into transcriptional regulation.* Nat Rev Genet, 2012. **13**(4): p. 233-45.

187. Tian, B., et al., *A large-scale analysis of mRNA polyadenylation of human and mouse genes.* Nucleic Acids Research, 2005. **33**(1): p. 201-212.

188. Batut, P., et al., *High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression.* Genome Res, 2013. **23**(1): p. 169-80.

189. Shepard, P.J., et al., *Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq.* RNA (New York, N.Y.), 2011. **17**(4): p. 761-772.

190. Shenker, S., et al., *IsoSCM: improved and alternative 3' UTR annotation using multiple change-point inference.* RNA, 2015. **21**(1): p. 14-27.

191. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature Biotechnology, 2015. **33**(3): p. 290-295.

192. Shao, M., J. Ma, and S. Wang, *DeepBound: accurate identification of transcript boundaries via deep convolutional neural fields.* Bioinformatics, 2017. **33**(14): p. i267-i273.

193. Garber, M., et al., *Computational methods for transcriptome annotation and quantification using RNA-seq.* Nature Methods, 2011. **8**(6): p. 469-477.

194. Guttman, M., et al., *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol, 2010. **28**(5): p. 503-10.

195. Huber, W., J. Toedling, and L.M. Steinmetz, *Transcript mapping with high-density oligonucleotide tiling arrays.* Bioinformatics, 2006. **22**(16): p. 1963-70.

196. Steijger, T., et al., *Assessment of transcript reconstruction methods for RNA-seq.* Nature Methods, 2013. **10**(12): p. 1177-1184.

197. Chen, M., et al., *A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data.* Brief Bioinform, 2019.

198. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nat Methods, 2010. **7**(12): p. 1009-15.

199. Ha, K.C.H., B.J. Blencowe, and Q. Morris, *QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data.* Genome Biol, 2018. **19**(1): p. 45.

200. Gruber, A.J., et al., *Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC.* Genome Biol, 2018. **19**(1): p. 44.

201. Birol, I., et al., *Kleat: cleavage site analysis of transcriptomes.* Pac Symp Biocomput, 2015: p. 347-58.

202. Bonfert, T. and C.C. Friedel, *Prediction of Poly(A) Sites by Poly(A) Read Mapping.* PLoS One, 2017. **12**(1): p. e0170914.

203. Szkop, K.J. and I. Nobeli, *Untranslated Parts of Genes Interpreted: Making Heads or Tails of High-Throughput Transcriptomic Data via Computational Methods: Computational methods to discover and quantify isoforms with alternative untranslated regions.* Bioessays, 2017. **39**(12).

204. Bayerlova, M., et al., *Newly Constructed Network Models of Different WNT Signaling Cascades Applied to Breast Cancer Expression Data.* PLoS One, 2015. **10**(12): p. e0144014.

205. Xia, Z., et al., *Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types.* Nature Communications, 2014. **5**: p. 5274.

206. Ye, C., et al., *APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data.* Bioinformatics, 2018. **34**(11): p. 1841-1849.

207. Arefeen, A., et al., *TAPAS: tool for alternative polyadenylation site analysis.* Bioinformatics, 2018. **34**(15): p. 2521-2529.

208. Derti, A., et al., *A quantitative atlas of polyadenylation in five mammals.* Genome Research, 2012. **22**(6): p. 1173-1183.

209. Hoque, M., et al., *Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing.* Nature Methods, 2013. **10**(2): p. 133-139.

210. Elkon, R., A.P. Ugalde, and R. Agami, *Alternative cleavage and polyadenylation: extent, regulation and function.* Nature Reviews Genetics, 2013. **14**(7): p. 496-506.

211. Ji, G., et al., *Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes.* Brief Bioinform, 2015. **16**(2): p. 304-13.

212. Gerard, C., *The basics of machine learning*, in *Practical Machine Learning in JavaScript: TensorFlow.js for Web Developers*, C. Gerard, Editor. 2021, Apress: Berkeley, CA. p. 1-24.

213. Bzdok, D., N. Altman, and M. Krzywinski, *Statistics versus machine learning.* Nat Methods, 2018. **15**(4): p. 233-234.

214. Libbrecht, M.W. and W.S. Noble, *Machine learning applications in genetics and genomics.* Nature Reviews Genetics, 2015. **16**(6): p. 321-332.

215. Cohen, S., *Chapter 2 - The basics of machine learning: strategies and techniques*, in *Artificial Intelligence and Deep Learning in Pathology*, S. Cohen, Editor. 2021, Elsevier. p. 13-40.

216. Li, W., et al. *A model based on eye movement data and artificial neutral network for product styling evaluation*. in *2018 24th International Conference on Automation and Computing (ICAC)*. 2018.

217. Aloysius, N. and M. Geetha. *A review on deep convolutional neural networks*. in *2017 International Conference on Communication and Signal Processing (ICCSP)*. 2017.

218. Lipton, Z.C., *A Critical Review of Recurrent Neural Networks for Sequence Learning.* ArXiv, 2015. **abs/1506.00019**.

219. Bengio, Y., P. Simard, and P. Frasconi, *Learning long-term dependencies with gradient descent is difficult.* IEEE Transactions on Neural Networks, 1994. **5**(2): p. 157-166.

220. Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural Comput, 1997. **9**(8): p. 1735-80.

221. Schuster, M. and K.K. Paliwal, *Bidirectional recurrent neural networks.* Ieee Transactions on Signal Processing, 1997. **45**(11): p. 2673-2681.

222. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.

223. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

224. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

225. GWAS Catalog, https://www.ebi.ac.uk/gwas/. Accessed 04 Dec. 2017.

226. Edenberg, H.J. and T. Foroud, *Genetics and alcoholism.* Nat Rev Gastroenterol Hepatol, 2013. **10**(8): p. 487-94.

227. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.

228. Perez-Enciso, M., J.R. Quevedo, and A. Bahamonde, *Genetical genomics: use all data.* BMC Genomics, 2007. **8**: p. 69.

229. Schadt, E.E., et al., *Genetics of gene expression surveyed in maize, mouse and man.* Nature, 2003. **422**(6929): p. 297-302.

230. Ghazalpour, A., et al., *Integrating genetic and network analysis to characterize genes related to mouse weight.* PLoS Genet, 2006. **2**(8): p. e130.

231. Mackay, T.F., E.A. Stone, and J.F. Ayroles, *The genetics of quantitative traits: challenges and prospects.* Nat Rev Genet, 2009. **10**(8): p. 565-77.

232. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules.* Science, 2003. **302**(5643): p. 249-55.

233. Ge, H., et al., *Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.* Nat Genet, 2001. **29**(4): p. 482-6.

234. Serin, E.A., et al., *Learning from Co-expression Networks: Possibilities and Challenges.* Front Plant Sci, 2016. **7**: p. 444.

235. Zhang, B. and S. Horvath, *A general framework for weighted gene co-expression network analysis.* Stat Appl Genet Mol Biol, 2005. **4**: p. Article17.

236. Oldham, M.C., et al., *Functional organization of the transcriptome in human brain.* Nat Neurosci, 2008. **11**(11): p. 1271-82.

237. Konopka, G., et al., *Human-specific transcriptional regulation of CNS development genes by FOXP2.* Nature, 2009. **462**(7270): p. 213-7.

238. DiLeo, M.V., et al., *Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome.* PLoS One, 2011. **6**(10): p. e26683.

239. Xue, J., et al., *Transcriptome-based network analysis reveals a spectrum model of human macrophage activation.* Immunity, 2014. **40**(2): p. 274-88.

240. Harrall, K.K., et al., *Uncovering the liver's role in immunity through RNA co-expression networks.* Mamm Genome, 2016. **27**(9-10): p. 469-84.

241. Bosron, W.F., T. Ehrig, and T.K. Li, *Genetic factors in alcohol metabolism and alcoholism.* Semin Liver Dis, 1993. **13**(2): p. 126-35.

242. Ramchandani, V.A., W.F. Bosron, and T.K. Li, *Research advances in ethanol metabolism.* Pathol Biol (Paris), 2001. **49**(9): p. 676-82.

243. Norberg, A., et al., *Role of variability in explaining ethanol pharmacokinetics: research and forensic applications.* Clin Pharmacokinet, 2003. **42**(1): p. 1-31.

244. Guindalini, C., et al., *Association of genetic variants in alcohol dehydrogenase 4 with alcohol dependence in Brazilian patients.* Am J Psychiatry, 2005. **162**(5): p. 1005-7.

245. Zakhari, S., *Overview: how is alcohol metabolized by the body?* Alcohol Res Health, 2006. **29**(4): p. 245-54.

246. Cederbaum, A.I., *Alcohol metabolism.* Clin Liver Dis, 2012. **16**(4): p. 667-85.

247. Grisel, J.E., et al., *Mapping of quantitative trait loci underlying ethanol metabolism in BXD recombinant inbred mouse strains.* Alcohol Clin Exp Res, 2002. **26**(5): p. 610-6.

248. Pravenec, M., et al., *An analysis of spontaneous hypertension in spontaneously hypertensive rats by means of new recombinant inbred strains.* J Hypertens, 1989. **7**(3): p. 217-21.

249. Elzhov, T.V., et al., *R interface to the Levenberg-Marquardt nonlinear least-squares algorithm found in MINPACK, plus support for bounds.* Retrieved from CRAN, 2012. **http://doi.org/10.2172/803290**.

250. Vaglenova, J., et al., *Expression, localization and potential physiological significance of alcohol dehydrogenase in the gastrointestinal tract.* Eur J Biochem, 2003. **270**(12): p. 2652-62.

251. Jaki, T. and M.J. Wolfsegger, *Estimation of pharmacokinetic parameters with the R package PK.* Pharmaceutical Statistics, 2010. **10**(3): p. 284-288.

252. Smit, A., R. Hubley, and P. Green, *RepeatMasker.* Open-3.0, 1996.

253. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002b. **12**(6): p. 996-1006.

254. UCSC Genome Browser. https://genome.ucsc.edu/. Accessed 15 April 2016.

255. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

256. Cunningham, F., et al., *Ensembl 2015.* Nucleic Acids Res, 2015. **43**(Database issue): p. D662-9.

257. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 323.

258. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.

259. Star Consortium, et al., *SNP and haplotype mapping for genetic analysis in the rat.* Nat Genet, 2008. **40**(5): p. 560-6.

260. PhenoGen. http://phenogen.ucdenver.edu. Accessed 04 Dec. 2017.

261. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nat Biotechnol, 2014. **32**(9): p. 896-902.

262. McCarthy, D.J., Y. Chen, and G.K. Smyth, *Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.* Nucleic Acids Res, 2012. **40**(10): p. 4288-97.

263. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

264. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

265. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.

266. Havlak, P., et al., *The Atlas genome assembly system.* Genome Res, 2004. **14**(4): p. 721-32.

267. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002a. **12**(4): p. 656-64.

268. Broman, K.W., et al., *R/qtl: QTL mapping in experimental crosses.* Bioinformatics, 2003. **19**(7): p. 889-90.

269. Churchill, G.A. and R.W. Doerge, *Empirical threshold values for quantitative trait mapping.* Genetics, 1994. **138**(3): p. 963-71.

270. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nat Genet, 1995. **11**(3): p. 241-7.

271. The Complex Trait Consortium, *Guidelines: The nature and identification of quantitative trait loci: a community's view.* Nature Reviews Genetics, 2003. **4**: p. 911-916.

272. Sen, S. and G.A. Churchill, *A statistical framework for quantitative trait mapping.* Genetics, 2001. **159**(1): p. 371-87.

273. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.

274. Aken, B.L., et al., *The Ensembl gene annotation system.* Database (Oxford), 2016. **2016**.

275. Hoog, J.O., et al., *Mammalian alcohol dehydrogenase - functional and structural implications.* J Biomed Sci, 2001. **8**(1): p. 71-6.

276. Zhang, Z., et al., *A greedy algorithm for aligning DNA sequences.* J Comput Biol, 2000. **7**(1-2): p. 203-14.

277. Hoog, J.O. and L.J. Ostberg, *Mammalian alcohol dehydrogenases--a comparative investigation at gene and protein levels.* Chem Biol Interact, 2011. **191**(1-3): p. 2-7.

278. Julia, P., J. Farres, and X. Pares, *Characterization of three isoenzymes of rat alcohol dehydrogenase. Tissue distribution and physical and enzymatic properties.* Eur J Biochem, 1987. **162**(1): p. 179-89.

279. Boleda, M.D., et al., *Role of extrahepatic alcohol dehydrogenase in rat ethanol metabolism.* Arch Biochem Biophys, 1989. **274**(1): p. 74-81.

280. Julia, P., X. Pares, and H. Jornvall, *Rat liver alcohol dehydrogenase of class III. Primary structure, functional consequences and relationships to other alcohol dehydrogenases.* Eur J Biochem, 1988. **172**(1): p. 73-83.

281. Estonius, M., et al., *Distribution of alcohol and sorbitol dehydrogenases. Assessment of mRNA species in mammalian tissues.* Eur J Biochem, 1993. **215**(2): p. 497-503.

282. Svensson, S., P. Stromberg, and J.O. Hoog, *A novel subtype of class II alcohol dehydrogenase in rodents. Unique Pro(47) and Ser(182) modulates hydride transfer in the mouse enzyme.* J Biol Chem, 1999. **274**(42): p. 29712-9.

283. Plapp, B.V., et al., *Contribution of liver alcohol dehydrogenase to metabolism of alcohols in rats.* Chem Biol Interact, 2015. **234**: p. 85-95.

284. Kardon, T., et al., *Identification of the gene encoding hydroxyacid-oxoacid transhydrogenase, an enzyme that metabolizes 4-hydroxybutyrate.* FEBS Lett, 2006. **580**(9): p. 2347-50.

285. Pares, X., et al., *Class IV alcohol dehydrogenase (the gastric enzyme). Structural analysis of human sigma sigma-ADH reveals class IV to be variable and confirms the presence of a fifth mammalian alcohol dehydrogenase class.* FEBS Lett, 1992. **303**(1): p. 69-72.

286. David, J.R., et al., *Biological role of alcohol dehydrogenase in the tolerance of Drosophila melanogaster to aliphatic alochols: utilization of an ADH-null mutant.* Biochem Genet, 1976. **14**(11-12): p. 989-97.

287. Ehrig, T., et al., *Degradation of aliphatic alcohols by human liver alcohol dehydrogenase: effect of ethanol and pharmacokinetic implications.* Alcohol Clin Exp Res, 1988. **12**(6): p. 789-94.

288. Boleda, M.D., et al., *Physiological substrates for rat alcohol dehydrogenase classes: aldehydes of lipid peroxidation, omega-hydroxyfatty acids, and retinoids.* Arch Biochem Biophys, 1993. **307**(1): p. 85-90.

289. Sellin, S., et al., *Oxidation and reduction of 4-hydroxyalkenals catalyzed by isozymes of human alcohol dehydrogenase.* Biochemistry, 1991. **30**(9): p. 2514-8.

290. Hartley, D.P., J.A. Ruth, and D.R. Petersen, *The hepatocellular metabolism of 4-hydroxynonenal by alcohol dehydrogenase, aldehyde dehydrogenase, and glutathione S-transferase.* Arch Biochem Biophys, 1995. **316**(1): p. 197-205.

291. Kumar, S., et al., *Alcohol and aldehyde dehydrogenases: retinoid metabolic effects in mouse knockout models.* Biochim Biophys Acta, 2012. **1821**(1): p. 198-205.

292. Shiota, G. and K. Kanki, *Retinoids and their target genes in liver functions and diseases.* J Gastroenterol Hepatol, 2013. **28 Suppl 1**: p. 33-7.

293. Matsumiya, T. and D.M. Stafforini, *Function and regulation of retinoic acid-inducible gene-I.* Crit Rev Immunol, 2010. **30**(6): p. 489-513.

294. Burd, C.G., T.I. Strochlic, and S.R. Setty, *Arf-like GTPases: not so Arf-like after all.* Trends Cell Biol, 2004. **14**(12): p. 687-94.

295. Britzen-Laurent, N., et al., *Intracellular trafficking of guanylate-binding proteins is regulated by heterodimerization in a hierarchical manner.* PLoS One, 2010. **5**(12): p. e14246.

296. Feng, J., et al., *Inducible GBP5 Mediates the Antiviral Response via Interferon-Related Pathways during Influenza A Virus Infection.* J Innate Immun, 2017. **9**(4): p. 419-435.

297. Hotter, D., et al., *Primate lentiviruses use at least three alternative strategies to suppress NF-kappaB-mediated immune activation.* PLoS Pathog, 2017. **13**(8): p. e1006598.

298. Xu, D., et al., *Mutational study of heparan sulfate 2-O-sulfotransferase and chondroitin sulfate 2-O-sulfotransferase.* J Biol Chem, 2007. **282**(11): p. 8356-67.

299. Schenauer, M.R., et al., *CCR2 chemokines bind selectively to acetylated heparan sulfate octasaccharides.* J Biol Chem, 2007. **282**(35): p. 25182-8.

300. Monneau, Y., F. Arenzana-Seisdedos, and H. Lortat-Jacob, *The sweet spot: how GAGs help chemokines guide migrating cells*. J Leukoc Biol, 2016. **99**(6): p. 935-53.

301. Lyon, M., G. Rushton, and J.T. Gallagher, *The interaction of the transforming growth factor-betas with heparin/heparan sulfate is isoform-specific.* J Biol Chem, 1997. **272**(29): p. 18000-6.

302. Khanna, M., et al., *Heparan sulfate as a receptor for poxvirus infections and as a target for antiviral agents.* J Gen Virol, 2017.

303. Kim, S.Y., et al., *Interaction of Zika Virus Envelope Protein with Glycosaminoglycans.* Biochemistry, 2017. **56**(8): p. 1151-1162.

304. Sasisekharan, R. and G. Venkataraman, *Heparin and heparan sulfate: biosynthesis, structure and function.* Curr Opin Chem Biol, 2000. **4**(6): p. 626-31.

305. Kreuger, J., et al., *Interactions between heparan sulfate and proteins: the concept of specificity.* J Cell Biol, 2006. **174**(3): p. 323-7.

306. Mishra-Gorur, K., H.A. Singer, and J.J. Castellot, Jr., *Heparin inhibits phosphorylation and autonomous activity of Ca(2+)/calmodulin-dependent protein kinase II in vascular smooth muscle cells.* Am J Pathol, 2002. **161**(5): p. 1893-901.

307. Mishra, J.P., et al., *Differential involvement of calmodulin-dependent protein kinase II-activated AP-1 and c-Jun N-terminal kinase-activated EGR-1 signaling pathways in*

*tumor necrosis factor-alpha and lipopolysaccharide-induced CD44 expression in human monocytic cells.* J Biol Chem, 2005. **280**(29): p. 26825-37.

308. Bailey, S.D., et al., *ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters.* Nat Commun, 2015. **2**: p. 6186.

309. Rizzo, F., et al., *Timed regulation of P-element-induced wimpy testis-interacting RNA expression during rat liver regeneration.* Hepatology, 2014. **60**(3): p. 798-806.

310. Taylor, A.L., et al., *Association of Hepatitis C Virus With Alcohol Use Among U.S. Adults: NHANES 2003-2010.* Am J Prev Med, 2016. **51**(2): p. 206-215.

311. McCartney, E.M., et al., *Alcohol metabolism increases the replication of hepatitis C virus and attenuates the antiviral action of interferon.* J Infect Dis, 2008. **198**(12): p. 1766-75.

312. Klyosov, A.A., et al., *Possible role of liver cytosolic and mitochondrial aldehyde dehydrogenases in acetaldehyde metabolism.* Biochemistry, 1996. **35**(14): p. 4445-56.

313. Wall, T.L., et al., *A genetic association with the development of alcohol and other substance use behavior in Asian Americans.* J Abnorm Psychol, 2001. **110**(1): p. 173-8.

314. Li, T.K., et al., *Genetic and environmental influences on alcohol metabolism in humans.* Alcohol Clin Exp Res, 2001. **25**(1): p. 136-44.

315. Luczak, S.E., S.J. Glatt, and T.J. Wall, *Meta-analyses of ALDH2 and ADH1B with alcohol dependence in Asians.* Psychological Bulletin, 2006. **132**(4): p. 607-621.

316. Wall, T.L., et al., *Alcohol metabolism in Asian-American men with genetic polymorphisms of aldehyde dehydrogenase.* Ann Intern Med, 1997. **127**(5): p. 376-9.

317. Chen, Y.C., et al., *Alcohol metabolism and cardiovascular response in an alcoholic patient homozygous for the ALDH2*2 variant gene allele.* Alcohol Clin Exp Res, 1999. **23**(12): p. 1853-60.

318. Luczak, S.E., et al., *ALDH2 status and conduct disorder mediate the relationship between ethnicity and alcohol dependence in Chinese, Korean, and White American college students.* J Abnorm Psychol, 2004. **113**(2): p. 271-8.

319. Wall, T.L., S.E. Luczak, and S. Hiller-Sturmhöfel, *Biology, Genetics, and Environment: Underlying Factors Influencing Alcohol Metabolism.* Alcohol research: current reviews, 2016. **38**(1): p. 59-68.

320. Eriksson, C.J. and H.W. Sippel, *The distribution and metabolism of acetaldehyde in rats during ethanol oxidation-I. The distribution of acetaldehyde in liver, brain, blood and breath.* Biochem Pharmacol, 1977. **26**(3): p. 241-7.

321. Hipolito, L., et al., *Brain metabolism of ethanol and alcoholism: an update.* Curr Drug Metab, 2007. **8**(7): p. 716-27.

322. Tabakoff, B., R.A. Anderson, and R.F. Ritzmann, *Brain acetaldehyde after ethanol administration.* Biochem Pharmacol, 1976. **25**(11): p. 1305-9.

323. Quertemont, E. and S. Tambour, *Is ethanol a pro-drug? The role of acetaldehyde in the central effects of ethanol.* Trends in Pharmacological Sciences, 2004. **25**(3): p. 130-134.

324. Peana, A.T., et al., *Mystic Acetaldehyde: The Never-Ending Story on Alcoholism.* Front Behav Neurosci, 2017. **11**: p. 81.

325. Aragon, C.M., F. Rogan, and Z. Amit, *Ethanol metabolism in rat brain homogenates by a catalase-H2O2 system.* Biochem Pharmacol, 1992. **44**(1): p. 93-8.

326. Zimatkin, S.M., A.V. Liopo, and R.A. Deitrich, *Distribution and kinetics of ethanol metabolism in rat brain.* Alcohol Clin Exp Res, 1998. **22**(8): p. 1623-7.

327. Gill, K., et al., *Enzymatic production of acetaldehyde from ethanol in rat brain tissue.* Alcohol Clin Exp Res, 1992. **16**(5): p. 910-5.

328. Aragon, C.M., L.M. Stotland, and Z. Amit, *Studies on ethanol-brain catalase interaction: evidence for central ethanol oxidation.* Alcohol Clin Exp Res, 1991. **15**(2): p. 165-9.

329. Aragon, C.M. and Z. Amit, *The effect of 3-amino-1,2,4-triazole on voluntary ethanol consumption: evidence for brain catalase involvement in the mechanism of action.* Neuropharmacology, 1992. **31**(7): p. 709-12.

330. Zimatkin, S.M., et al., *Enzymatic mechanisms of ethanol oxidation in the brain.* Alcohol Clin Exp Res, 2006. **30**(9): p. 1500-5.

331. Del Maestro, R. and W. McDonald, *Distribution of superoxide dismutase, glutathione peroxidase and catalase in developing rat brain.* Mech Ageing Dev, 1987. **41**(1-2): p. 29-38.

332. Cohen, G., P.M. Sinet, and R. Heikkila, *Ethanol oxidation by rat brain in vivo.* Alcohol Clin Exp Res, 1980. **4**(4): p. 366-70.

333. Peana, A.T., et al., *Key role of ethanol-derived acetaldehyde in the motivational properties induced by intragastric ethanol: a conditioned place preference study in the rat.* Alcohol Clin Exp Res, 2008. **32**(2): p. 249-58.

334. Font, L., M. Miquel, and C.M. Aragon, *Involvement of brain catalase activity in the acquisition of ethanol-induced conditioned place preference.* Physiol Behav, 2008. **93**(4-5): p. 733-41.

335. Peana, A.T., et al., *l-Cysteine reduces oral ethanol self-administration and reinstatement of ethanol-drinking behavior in rats.* Pharmacol Biochem Behav, 2010. **94**(3): p. 431-7.

336. March, S.M., et al., *The role of acetaldehyde in ethanol reinforcement assessed by Pavlovian conditioning in newborn rats.* Psychopharmacology (Berl), 2013. **226**(3): p. 491-9.

337. Font, L., M. Miquel, and C.M. Aragon, *Prevention of ethanol-induced behavioral stimulation by D-penicillamine: a sequestration agent for acetaldehyde.* Alcohol Clin Exp Res, 2005. **29**(7): p. 1156-64.

338. Font, L., C.M. Aragon, and M. Miquel, *Voluntary ethanol consumption decreases after the inactivation of central acetaldehyde by d-penicillamine.* Behav Brain Res, 2006. **171**(1): p. 78-86.

339. Vinci, S., et al., *Acetaldehyde elicits ERK phosphorylation in the rat nucleus accumbens and extended amygdala.* Synapse, 2010. **64**(12): p. 916-27.

340. Peana, A.T., G. Muggironi, and M. Diana, *Acetaldehyde-reinforcing effects: a study on oral self-administration behavior.* Front Psychiatry, 2010. **1**: p. 23.

341. Orrico, A., et al., *Efficacy of D-penicillamine, a sequestering acetaldehyde agent, in the prevention of alcohol relapse-like drinking in rats.* Psychopharmacology (Berl), 2013. **228**(4): p. 563-75.

342. Amit, Z., Z.W. Brown, and G.E. Rockman, *Possible involvement of acetaldehyde, norepinephrine and their tetrahydroisoquinoline derivatives in the regulation of ethanol seld-administration.* Drug Alcohol Depend, 1977. **2**(5-6): p. 495-500.

343. Brown, Z.W., Z. Amit, and B. Smith, *Intraventricular self-administration of acetaldehyde and voluntary consumption of ethanol in rats.* Behav Neural Biol, 1980. **28**(2): p. 150-5.

344. Hernandez, J.A., R.C. Lopez-Sanchez, and A. Rendon-Ramirez, *Lipids and Oxidative Stress Associated with Ethanol-Induced Neurological Damage.* Oxid Med Cell Longev, 2016. **2016**: p. 1543809.

345. Puig, J.G. and I.H. Fox, *Ethanol-induced activation of adenine nucleotide turnover. Evidence for a role of acetate.* J Clin Invest, 1984. **74**(3): p. 936-41.

346. Sarkola, T., et al., *Ethanol, acetaldehyde, acetate, and lactate levels after alcohol intake in white men and women: effect of 4-methylpyrazole.* Alcohol Clin Exp Res, 2002. **26**(2): p. 239-45.

347. Correa, M., et al., *Piecing together the puzzle of acetaldehyde as a neuroactive agent.* Neuroscience & Biobehavioral Reviews, 2012. **36**(1): p. 404-430.

348. Soliman, M.L. and T.A. Rosenberger, *Acetate supplementation increases brain histone acetylation and inhibits histone deacetylase activity and expression.* Mol Cell Biochem, 2011. **352**(1-2): p. 173-80.

349. Moonat, S., et al., *Aberrant histone deacetylase2-mediated histone modifications and synaptic plasticity in the amygdala predisposes to anxiety and alcoholism.* Biol Psychiatry, 2013. **73**(8): p. 763-73.

350. Sakharkar, A.J., et al., *Effects of histone deacetylase inhibitors on amygdaloid histone acetylation and neuropeptide Y expression: a role in anxiety-like and alcohol-drinking behaviours.* Int J Neuropsychopharmacol, 2014. **17**(8): p. 1207-20.

351. Li, T.K., L. Lumeng, and D.P. Doolittle, *Selective breeding for alcohol preference and associated responses.* Behav Genet, 1993. **23**(2): p. 163-70.

352. Stewart, R.B., et al., *Comparison of alcohol-preferring (P) and nonpreferring (NP) rats on tests of anxiety and for the anxiolytic effects of ethanol.* Alcohol, 1993. **10**(1): p. 1-10.

353. Pandey, S.C., et al., *Deficits in amygdaloid cAMP-responsive element-binding protein signaling play a role in genetic predisposition to anxiety and alcoholism.* J Clin Invest, 2005. **115**(10): p. 2762-73.

354. Mews, P., et al., *Alcohol metabolism contributes to brain histone acetylation.* Nature, 2019. **574**(7780): p. 717-721.

355. Mulligan, M.K., et al., *Molecular profiles of drinking alcohol to intoxication in C57BL/6J mice.* Alcohol Clin Exp Res, 2011. **35**(4): p. 659-70.

356. Volkow, N.D., et al., *Acute alcohol intoxication decreases glucose metabolism but increases acetate uptake in the human brain.* Neuroimage, 2013. **64**: p. 277-83.

357. Pawlosky, R.J., et al., *Alterations in brain glucose utilization accompanying elevations in blood ethanol and acetate concentrations in the rat.* Alcohol Clin Exp Res, 2010. **34**(2): p. 375-81.

358. Volkow, N.D., et al., *Acute effects of ethanol on regional brain glucose metabolism and transport.* Psychiatry Res, 1990. **35**(1): p. 39-48.

359. Nuutinen, H., et al., *Elevated blood acetate as indicator of fast ethanol elimination in chronic alcoholics.* Alcohol, 1985. **2**(4): p. 623-6.

360. Jiang, L., et al., *Increased brain uptake and oxidation of acetate in heavy drinkers.* J Clin Invest, 2013. **123**(4): p. 1605-14.

361. Enculescu, C., et al., *Proteomics Reveals Profound Metabolic Changes in the Alcohol Use Disorder Brain.* ACS Chem Neurosci, 2019. **10**(5): p. 2364-2373.

362. Nair, K.P., *Alcohol, Brain Function, and Behavioral Impact*, in *Food and Human Responses : A Holistic View*. 2020, Springer International Publishing: Cham. p. 143-154.

363. Derr, R.F., K. Draves, and M. Derr, *Abatement by acetate of an ethanol withdrawal syndrome.* Life Sci, 1981. **29**(17): p. 1787-90.

364. Park, J.Y., et al., *Comparative analysis of mRNA isoform expression in cardiac hypertrophy and development reveals multiple post-transcriptional regulatory modules.* PLoS ONE, 2011. **6**(7): p. e22391.

365. de Klerk, E., et al., *Poly(A) binding protein nuclear 1 levels affect alternative polyadenylation.* Nucleic Acids Research, 2012. **40**(18): p. 9089-9101.

366. Jenal, M., et al., *The Poly(A)-Binding Protein Nuclear 1 Suppresses Alternative Cleavage and Polyadenylation Sites.* Cell, 2012. **149**(3): p. 538-553.

367. Lembo, A., F. Di Cunto, and P. Provero, *Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer.* PLoS One, 2012. **7**(2): p. e31129.

368. Bishop, D.F., R. Kornreich, and R.J. Desnick, *Structural organization of the human alpha-galactosidase A gene: further evidence for the absence of a 3' untranslated region.* Proceedings of the National Academy of Sciences, 1988. **85**(11): p. 3903-3907.

369. Lin, C.L., et al., *Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis.* Neuron, 1998. **20**(3): p. 589-602.

370. Gieselmann, V., et al., *Arylsulfatase A pseudodeficiency: loss of a polyadenylylation signal and N-glycosylation site.* Proceedings of the National Academy of Sciences, 1989. **86**(23): p. 9436-9440.

371. Lemmers, R.J.L.F., et al., *A unifying genetic model for facioscapulohumeral muscular dystrophy.* Science, 2010. **329**(5999): p. 1650-1653.

372. Schurch, N.J., et al., *Improved Annotation of 3' Untranslated Regions and Complex Loci by Combination of Strand-Specific Direct RNA Sequencing, RNA-Seq and ESTs.* PLOS ONE, 2014. **9**(4): p. e94270.

373. Consortium, M., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nature Biotechnology, 2006. **24**(9): p. 1151-1161.

374. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

375. Palomares, M.A., et al., *Systematic analysis of TruSeq, SMARTer and SMARTer Ultra-Low RNA-seq kits for standard, low and ultra-low quantity samples.* Sci Rep, 2019. **9**(1): p. 7550.

376. Schuierer, S., et al., *A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples.* BMC Genomics, 2017. **18**(1): p. 442.

377. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10.

378. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nat Biotechnol, 2019. **37**(8): p. 907-915.

379. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

380. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Res, 2004. **32**(Database issue): p. D493-6.

381. Kent, W.J., et al., *The Human Genome Browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

382. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D682-D688.

383. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

384. Miura, P., et al., *Alternative polyadenylation in the nervous system: to what lengths will 3' UTR extensions take us?* BioEssays : news and reviews in molecular, cellular and developmental biology, 2014. **36**(8): p. 766-777.

385. Neve, J., et al., *Cleavage and polyadenylation: Ending the message expands gene regulation.* RNA Biol, 2017. **14**(7): p. 865-890.

386. Beaudoing, E., et al., *Patterns of variant polyadenylation signal usage in human genes.* Genome Research, 2000. **10**(7): p. 1001-1010.

387. Hu, J., et al., *Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation.* RNA, 2005. **11**(10): p. 1485-93.

388. Salisbury, J., K.W. Hutchison, and J.H. Graber, *A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-64 RNA recognition motif.* BMC Genomics, 2006. **7**: p. 55.

389. Hutchins, L.N., et al., *Position-dependent motif characterization using non-negative matrix factorization.* Bioinformatics, 2008. **24**(23): p. 2684-90.

390. Legendre, M. and D. Gautheret, *Sequence determinants in human polyadenylation site selection.* BMC Genomics, 2003. **4**(1): p. 7.

391. McDevitt, M.A., et al., *Sequences capable of restoring poly(A) site function define two distinct downstream elements.* EMBO J, 1986. **5**(11): p. 2907-13.

392. Gil, A. and N.J. Proudfoot, *Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation.* Cell, 1987. **49**(3): p. 399-406.

393. Kalkatawi, M., et al., *Dragon PolyA Spotter: predictor of poly(A) motifs within human genomic DNA sequences.* Bioinformatics, 2013. **29**(11): p. 1484.

394. Akhtar, M.N., et al., *POLYAR, a new computer program for prediction of poly(A) sites in human sequences.* BMC Genomics, 2010. **11**(1): p. 646.

395. Gers, F.A., J. Schmidhuber, and F. Cummins, *Learning to forget: continual prediction with LSTM.* Neural Comput, 2000. **12**(10): p. 2451-71.

396. Baldi, P., et al., *Exploiting the past and the future in protein secondary structure prediction.* Bioinformatics, 1999. **15**(11): p. 937-46.

397. Kingma, D.P. and J. Ba *Adam: A Method for Stochastic Optimization*. arXiv e-prints, 2014. arXiv:1412.6980.

398. Saito, T. and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets.* PLoS One, 2015. **10**(3): p. e0118432.

399. Bogard, N., et al., *A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation.* Cell, 2019. **178**(1): p. 91-106 e23.

400. Masamha, C.P., et al., *CFIm25 links alternative polyadenylation to glioblastoma tumour suppression.* Nature, 2014. **510**(7505): p. 412-6.

401. Li, B., et al., *A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq.* Sci Rep, 2017. **7**(1): p. 4200.

402. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

403. Ryan Lusk*, E.S., Farnoush Banaei-Kashani, Boris Tabakoff, Katerina Kechris, and Laura M. Saba, Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high throughput RNA sequencing and DNA sequence.

404. Herrmann, C.J., et al., *PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing.* Nucleic Acids Res, 2020. **48**(D1): p. D174-D179.

405. Wang, R., et al., *PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes.* Nucleic Acids Res, 2018. **46**(D1): p. D315-D319.

406. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.

407. Kubo, T., et al., *Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3'-UTRs.* Nucleic Acids Research, 2006. **34**(21): p. 6264-6271.

408. Pamplona, F.A., L.F. Vendruscolo, and R.N. Takahashi, *Increased sensitivity to cocaine-induced analgesia in Spontaneously Hypertensive Rats (SHR).* Behav Brain Funct, 2007. **3**: p. 9.

409. Vendruscolo, L.F., G.S. Izidio, and R.N. Takahashi, *Drug reinforcement in a rat model of attention deficit/hyperactivity disorder--the Spontaneously Hypertensive Rat (SHR).* Curr Drug Abuse Rev, 2009. **2**(2): p. 177-83.

410. Papadimitriou, E., et al., *Pleiotrophin and its receptor protein tyrosine phosphatase beta/zeta as regulators of angiogenesis and cancer.* Biochim Biophys Acta, 2016. **1866**(2): p. 252-265.

411. Le Grevès, P., *Pleiotrophin gene transcription in the rat nucleus accumbens is stimulated by an acute dose of amphetamine.* Brain Research Bulletin, 2005. **65**(6): p. 529-532.

412. Nilsen, T.W. and B.R. Graveley, *Expansion of the eukaryotic proteome by alternative splicing.* Nature, 2010. **463**(7280): p. 457-63.

413. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.

414. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

415. Hoque, M., et al., *Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing.* nature.com, 2013.

416. Garcia-Blanco, M.A., A.P. Baraniak, and E.L. Lasda, *Alternative splicing in disease and therapy.* Nature Biotechnology, 2004. **22**(5): p. 535-546.

417. Kelemen, O., et al., *Function of alternative splicing.* Gene, 2013. **514**(1): p. 1-30.

418. Ren, F., et al., *Alternative Polyadenylation: a new frontier in post transcriptional regulation.* Biomark Res, 2020. **8**(1): p. 67.

419. Wang, G.S. and T.A. Cooper, *Splicing in disease: disruption of the splicing code and the decoding machinery.* Nat Rev Genet, 2007. **8**(10): p. 749-61.

420. Nissim-Rafinia, M. and B. Kerem, *Splicing regulation as a potential genetic modifier.* Trends Genet, 2002. **18**(3): p. 123-7.

421. Hutton, M., et al., *Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17.* Nature, 1998. **393**(6686): p. 702-5.

422. Hong, M., et al., *Mutation-specific functional impairments in distinct tau isoforms of hereditary FTDP-17.* Science, 1998. **282**(5395): p. 1914-7.

423. Spillantini, M.G., et al., *Mutation in the tau gene in familial multiple system tauopathy with presenile dementia.* Proc Natl Acad Sci U S A, 1998. **95**(13): p. 7737-41.

424. Dredge, B.K., A.D. Polydorides, and R.B. Darnell, *The splice of life: alternative splicing and neurological disease.* Nat Rev Neurosci, 2001. **2**(1): p. 43-50.

425. Lusk, R., et al., *Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence.* Nat Commun, 2021. **12**(1): p. 1652.

426. Vanderlinden, L.A., et al., *Is the Alcohol Deprivation Effect Genetically Mediated? Studies with HXB/BXH Recombinant Inbred Rat Strains.* Alcoholism: Clinical and Experimental Research, 2014. **38**(7): p. 2148-2157.

427. Smit, A., R. Hubley, and P. Green, *RepeatMasker Open-3.0*. 1996.

428. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**(4): p. 357-359.

429. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nature Methods, 2015. **12**(4): p. 357-360.

430. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.

431. Saar, K., et al., *SNP and haplotype mapping for genetic analysis in the rat.* Nature Genetics, 2008. **40**(5): p. 560-566.

432. Yates, A.D., et al., *Ensembl 2020.* Nucleic Acids Research, 2020. **48**(D1): p. D682-D688.

433. Havlak, P., *The Atlas Genome Assembly System.* Genome Research, 2004. **14**(4): p. 721-732.

434. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC bioinformatics, 2010. **11**(1): p. 94.

435. Risso, D., et al., *GC-content normalization for RNA-Seq data.* BMC Bioinformatics, 2011. **12**: p. 480.

436. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* 2014.

437. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

438. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-3.

439. Kent, W.J., *BLAT---The BLAST-Like Alignment Tool.* Genome Research, 2002. **12**(4): p. 656-664.

440. Lusk, R., et al., *Unsupervised, statistically-based systems biology approach for unraveling the genetics of complex traits: A demonstration with ethanol metabolism.* Alcoholism: Clinical and Experimental Research, 2018.

441. Churchill, G.A. and R.W. Doerge, *Empirical threshold values for quantitative trait mapping.* Genetics, 1994. **138**(3): p. 963-971.

442. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nature Genetics, 1995. **11**(3): p. 241-247.

443. Abiola, O., et al., *The nature and identification of quantitative trait loci: a community's view.* Nat Rev Genet, 2003. **4**(11): p. 911-6.

444. Broman, K.W., et al., *R/qtl: QTL mapping in experimental crosses.* Bioinformatics, 2003. **19**(7): p. 889-890.

445. Saba, L.M., et al., *A long non-coding RNA (Lrap) modulates brain gene expression and levels of alcohol consumption in rats.* Genes Brain Behav, 2021. **20**(2): p. e12698.

446. Shannon, P., *Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.* Genome Research, 2003. **13**(11): p. 2498-2504.

447. Ji, X., et al., *A comprehensive rat transcriptome built from large scale RNA-seq-based annotation.* Nucleic Acids Res, 2020. **48**(15): p. 8320-8331.

448. MacDonald, C.C. and K.W. McMahon, *Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond.* Wiley Interdiscip Rev RNA, 2010. **1**(3): p. 494-501.

449. Cargnello, M. and P.P. Roux, *Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases.* Microbiol Mol Biol Rev, 2011. **75**(1): p. 50-83.

450. New, L., et al., *PRAK, a novel protein kinase regulated by the p38 MAP kinase.* EMBO J, 1998. **17**(12): p. 3372-84.

451. Lo, U., et al., *p38alpha (MAPK14) critically regulates the immunological response and the production of specific cytokines and chemokines in astrocytes.* Sci Rep, 2014. **4**: p. 7405.

452. Arlinde, C., et al., *A cluster of differentially expressed signal transduction genes identified by microarray analysis in a rat genetic model of alcoholism.* Pharmacogenomics J, 2004. **4**(3): p. 208-18.

453. Sommer, W., C. Arlinde, and M. Heilig, *The search for candidate genes of alcoholism: evidence from expression profiling studies.* Addict Biol, 2005. **10**(1): p. 71-9.

454. Rodd, Z.A., et al., *Candidate genes, pathways and mechanisms for alcoholism: an expanded convergent functional genomics approach.* Pharmacogenomics J, 2007. **7**(4): p. 222-56.

455. Martin-Blanco, E., *p38 MAPK signalling cascades: ancient roles and new functions.* Bioessays, 2000. **22**(7): p. 637-45.

456. Baik, I., et al., *Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men.* Am J Clin Nutr, 2011. **93**(4): p. 809-16.

457. Lee, W.B., et al., *OAS1 and OAS3 negatively regulate the expression of chemokines and interferon-responsive genes in human macrophages.* BMB Rep, 2019. **52**(2): p. 133-138.

458. DePaula-Silva, A.B., et al., *Differential transcriptional profiles identify microglial- and macrophage-specific gene markers expressed during virus-induced neuroinflammation.* J Neuroinflammation, 2019. **16**(1): p. 152.

459. Vanin, E.F., *Processed pseudogenes: characteristics and evolution.* Annu Rev Genet, 1985. **19**: p. 253-72.

460. Kalyana-Sundaram, S., et al., *Expressed pseudogenes in the transcriptional landscape of human cancers.* Cell, 2012. **149**(7): p. 1622-34.

461. Martinez, N.M., et al., *Alternative splicing networks regulated by signaling in human T cells.* RNA, 2012. **18**(5): p. 1029-40.

462. Martinez, N.M. and K.W. Lynch, *Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn.* Immunol Rev, 2013. **253**(1): p. 216-36.

463. McClintick, J.N., et al., *Gene expression changes in the ventral hippocampus and medial prefrontal cortex of adolescent alcohol-preferring (P) rats following binge-like alcohol drinking.* Alcohol, 2018. **68**: p. 37-47.

464. Black, W.J., et al., *Human aldehyde dehydrogenase genes: alternatively spliced transcriptional variants and their suggested nomenclature.* Pharmacogenet Genomics, 2009. **19**(11): p. 893-902.

465. Touloupi, K., et al., *The Basis for Strain-Dependent Rat Aldehyde Dehydrogenase 1A7 (ALDH1A7) Gene Expression.* Mol Pharmacol, 2019. **96**(5): p. 655-663.

466. Singh, S., et al., *Aldehyde dehydrogenases in cellular responses to oxidative/electrophilic stress.* Free Radic Biol Med, 2013. **56**: p. 89-101.

467. Oliveira, L.M., F.M.E. Teixeira, and M.N. Sato, *Impact of Retinoic Acid on Immune Cells and Inflammatory Diseases.* Mediators Inflamm, 2018. **2018**: p. 3067126.

468. Langfelder, P., et al., *A systems genetic analysis of high density lipoprotein metabolism and network preservation across mouse models.* Biochim Biophys Acta, 2012. **1821**(3): p. 435-47.

469. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biol, 2017. **18**(1): p. 83.

470. Rachamin, G., et al., *Modulation of alcohol dehydrogenase and ethanol metabolism by sex hormones in the spontaneously hypertensive rat. Effect of chronic ethanol administration.* The Biochemical journal, 1980. **186**(2): p. 483-490.

471. Fan, H., et al., *Comparison of Long Short Term Memory Networks and the Hydrological Model in Runoff Simulation.* Water, 2020. **12**(1).

472. Srivastava, N., et al., *Dropout: a simple way to prevent neural networks from overfitting.* J. Mach. Learn. Res., 2014. **15**(1): p. 1929–1958.

# APPENDIX A

# CHAPTER II SUPPLEMENTARY

## Supplementary Methods

*Detailed Description of Ethanol Delivery Method and Dose Choice Rationale*

The use of intraperitoneal delivery circumvents the gastrointestinal tract (GI), and therefore also any confounding influence of GI alcohol metabolism on alcohol clearance [1], while still otherwise closely mimicking first pass metabolism and the traditional route of alcohol administration [2, 3]. Indeed, the relative contribution of the GI tract to first pass metabolism of alcohol remains controversial; however, here we are only interested in the genetic variables of the main alcohol metabolizing organ – the liver [4, 5]. The dose of alcohol was chosen based on previous observations in these animals that it induces moderate intoxication and anxiolysis and achieves peak blood alcohol concentrations of approximately 40 mM. The high blood alcohol concentrations are presumably sufficient to saturate the alcohol metabolizing enzyme class I alcohol dehydrogenase 1 (ADH1), as well as capture any genetic involvement of low affinity, i.e. high $K_m$, enzyme systems on alcohol clearance such as the class II alcohol dehydrogenase 4 (ADH4) identified in the rat [6, 7].

*Detailed Description of Alcohol Clearance Quantitation in the HXB/BXH Recombinant Inbred Rat Panel*

For gas chromatographic analysis of blood alcohol levels, ambient headspace injections with an injection volume of 10 μL were performed, and nitrogen was utilized for the mobile phase. The 2-propanol (product number 675431) and the ethanol (product number 493511) used for the internal standards and standard curves, respectively, were received from Sigma-Aldrich (Sigma-Aldrich, St. Louis, MO, USA). Estimated concentrations below 2 mM were considered

under the detection limit and consequently removed from alcohol clearance calculations. Additionally, individual rats missing more than half (5 out of 10 time points) of their blood samples were removed from the analysis, and measurements at 0 minutes were not used in the model fitting as the pharmacokinetic (PK) model used assumed zero concentration at time 0.

*Detailed Description of Acetate Area Under the Curve Quantitation in the HXB/BXH Recombinant Inbred Rat Panel*

In general, each plate contained 30 blood samples in duplicate (60 measures total) and two standard curves; one was measured prior to the experimental samples and one measured after the experimental samples were recorded. The standard curves were first evaluated for quality using the coefficient of determination from a linear regression model predicting observed quantity from known acetate concentration. When both curves had coefficients of determination greater than 0.9, the average intercept and average slope was used for estimating acetate concentrations. When only one curve met this criterion, only that curve was used for quantitation. When neither curve had a coefficient of determination greater than 0.9, individual values were examined for quality. If the rank of two observed values did not match the rank of their known concentration, these two values were removed and the calibration curve was estimated based on four concentrations. If the rank of more than two observed values differed from their known rank, the observation with the largest absolute deviation in the fitted linear model was removed. Additionally, entire plates/batches that had greater than two-thirds of its absorbance measurements below background absorbance, defined as the mean absorbance of 0.00 M acetate concentration standards for the given batch, were not included. To ensure comparable acetate AUC estimates across rats, rats missing acetate values at the beginning (0 minutes) and/or end (400 minutes) of the acetate concentration-time curve were excluded from

further analysis. Acetate measurements were repeated for several rats excluded due to quality control using a separate aliquot of the blood samples and four standard curves per plate, and the resulting measurements were processed as noted above. Duplicate acetate measurements were averaged to give the concentration for a given sample. Negative acetate concentrations were counted as zero when calculating acetate AUC. Of note, adjusting for basal acetate levels, i.e. shifting the acetate concentration-time curves for each rat based on the concentration of acetate at 0 minutes when calculating acetate AUC, did not alter the conclusions of this study as most of these values were below the limits of detection.

*Detailed Description of Trim Galore!, TopHat, and RSEM Settings Used in the Quantitation of Whole Liver RNA Sequencing for the HXB/BXH Recombinant Inbred Rat Panel Pipeline*

Trim Galore!:

Function = trim_galore

Options = --paired –stringency 3 -q 20

TopHat:

Function = tophat

Options = -g 2 --library-type fr-firststrand -p 12

RSEM:

Function = rsem-calculate-expression

Options = -p 6 –time --seed-length 20 --bowtie2 --no-bamoutput --forward-prob=0.0 --paired-end

## References


1. T.-K. Li and T. K. Li, "The Absorption, Distribution, and Metabolism of Ethanol and Its Effects on Nutrition and Hepatic Function," in *Medical and Social Aspects of Alcohol Abuse*, no. 3, Boston, MA: Springer US, 1983, pp. 47–77.

2. A. Adam, J. Van Cantfort, and J. Gielen, "Kinetics of absorption, equilibration (or distribution), and excretion of orally and intraperitoneally administered cholesterol in the rat," *Journal of Pharmacology and …*, vol. 11, no. 8, pp. 610–615, Aug. 1976.

3. P. V. Turner, T. Brabb, C. Pekow, and M. A. Vasbinder, "Administration of Substances to Laboratory Animals: Routes of Administration and Factors to Consider," 2011.

4. S. Zakhari, "Overview: how is alcohol metabolized by the body?," *Alcohol Research & Health*, 2006.

5. A. I. Cederbaum, "Alcohol metabolism.," *Clin Liver Dis*, vol. 16, no. 4, pp. 667–685, Nov. 2012.

6. P. Julià, P. JULIA, J. FARRES, X. PARES, J. Farrés, and X. Parés, "Characterization of three isoenzymes of rat alcohol dehydrogenase. Tissue distribution and physical and enzymatic properties," *Eur. J. Biochem.*, vol. 162, no. 1, pp. 179–189, Jan. 1987.

7. M. D. Boleda, M. D. Boleda, P. JULIA, A. Moreno, P. Julià, X. PARES, A. Moreno, and X. Parés, "Role of extrahepatic alcohol dehydrogenase in rat ethanol metabolism," *Archives of biochemistry and …*, vol. 274, no. 1, pp. 74–81, Oct. 1989.


## Supplementary Results

*Results of Quality Control Imposed on Alcohol and Acetate Measures in the HXB/BXH Recombinant Inbred Rat Panel*

Of the 725 alcohol concentration measures (i.e., not including measures for 0 minutes), 33 were below 2 mM and removed prior to clearance calculations. One individual rat had only a single alcohol concentration estimate and therefore this sample was entirely removed from the analysis. Originally, acetate measurements were derived from 56 batches. However, six batches were removed due to large numbers of samples undetectable above background absorbance (five) or no data for acetate measures at 0 minutes (one). To replenish the lost data, nine

additional batches were added that consisted of the eliminated (or otherwise missing) rat blood samples for which there were extra stored blood aliquots; thus, acetate measurements came from a total of 59 batches. Of these, 45 used two standard curves to calculate acetate concentrations, five used one, eight used four, and one used three. Furthermore, 17 individual standards were identified as outliers and removed from standard curve calculations.

*Detailed Results From STAR Molecular Maker Set Processing*

19,391 of the 20,283 SNPs originally in the STAR uniquely and perfectly aligned to the Rnor_6.0 rat genome. Removing markers that did not differ between parental strains (BN-Lx/Cub and SHR/OlaIpcv) resulted in 13,103 remaining SNPs. SNPs were also removed if they were heterozygous in either progenitor strain or if they were heterozygous or missing for more than 5% of the HXB/BXH RI strains. This left 10,582 markers. Double recombinant and improbable recombination marker removal resulted in a final marker set of 10,486 SNPs for QTL analyses. Finally, 1,529 unique strain distribution patterns, i.e. haplotype blocks, were identified for the 32 HXB/BXH RI strains genotyped by the STAR Consortium.

## Supplementary Tables

### Supplementary Table S1. Statistical comparison of zero-order and first-order elimination phase pharmacokinetic model fits in individual animals of the HXB/BXH recombinant inbred rat panel.

| Animal | Strain | First-order Elimination Rate Constant (/min) | Sum of Residuals from First-order Model Fit | Zero-order Elimination Rate Constant (mM/min) | Sum of Residuals from Zero-order Model Fit | F-statistic (First-order Sum of Residuals/Zero-order Sum of Residuals) | p-value |
|---|---|---|---|---|---|---|---|
| BXH10_1_2G | BXH10 | 0.0023 | 154.3 | 0.0830 | 152.2 | 1.01 | 0.51 |
| BXH10_2_2G | BXH10 | 0.0013 | 51.3 | 0.0406 | 49.0 | 1.05 | 0.52 |
| BXH10_3_2G | BXH10 | 0.0021 | 78.7 | 0.0677 | 75.6 | 1.04 | 0.52 |
| BXH11-2mg-1 | BXH11 | 0.0025 | 69.4 | 0.0866 | 132.3 | 0.52 | 0.23 |
| BXH11-2mg-2 | BXH11 | 0.0026 | 118.4 | 0.0902 | 114.8 | 1.03 | 0.51 |
| BXH11-2mg-3 | BXH11 | 0.0030 | 35.6 | 0.0965 | 24.7 | 1.44 | 0.67 |
| BXH12-2mg-1 | BXH12 | 0.0018 | 99.3 | 0.0808 | 79.3 | 1.25 | 0.60 |
| BXH12-2mg-2 | BXH12 | 0.0023 | 50.4 | 0.0914 | 40.8 | 1.23 | 0.58 |
| BXH12-2mg-3 | BXH12 | 0.0022 | 109.2 | 0.0816 | 94.6 | 1.15 | 0.55 |
| BXH13-2mg-1 | BXH13 | 0.0038 | 1.0 | 0.0649 | 0.2 | 5.25 | 0.90 |
| BXH13-2mg-2 | BXH13 | 0.0035 | 241.3 | 0.1054 | 240.4 | 1.00 | 0.50 |
| BXH13-2mg-3 | BXH13 | 0.0055 | 79.5 | 0.1122 | 65.6 | 1.21 | 0.59 |
| BXH2-2mg-1 | BXH2 | 0.0024 | 118.2 | 0.0900 | 162.5 | 0.73 | 0.40 |
| BXH2-2mg-2 | BXH2 | 0.0032 | 38.0 | 0.0981 | 55.6 | 0.68 | 0.34 |
| BXH2-2mg-3 | BXH2 | 0.0023 | 287.0 | 0.1207 | 190.6 | 1.51 | 0.70 |
| BXH3_1_2G | BXH3 | 0.0020 | 33.5 | 0.0532 | 33.2 | 1.01 | 0.50 |
| BXH5-2mg-2 | BXH5 | 0.0023 | 19.2 | 0.0808 | 25.2 | 0.76 | 0.40 |
| BXH5-2mg-3 | BXH5 | 0.0029 | 182.5 | 0.1106 | 281.4 | 0.65 | 0.29 |
| BXH5-2mg-4 | BXH5 | 0.0023 | 217.8 | 0.0799 | 268.7 | 0.81 | 0.39 |
| BHX6_3_2G | BXH6 | 0.0056 | 79.8 | 0.1163 | 36.6 | 2.18 | 0.79 |
| BXH6_1_2G | BXH6 | 0.0065 | 72.7 | 0.1758 | 39.8 | 1.83 | 0.65 |
| BXH6_2_2G | BXH6 | 0.0056 | 52.0 | 0.1786 | 19.0 | 2.73 | 0.78 |
| BXH8_1_2G | BXH8 | 0.0066 | 162.8 | 0.2044 | 56.1 | 2.90 | 0.84 |
| BXH8_2_2G | BXH8 | 0.0057 | 95.0 | 0.1423 | 47.7 | 1.99 | 0.77 |
| BXH9_2_2G | BXH9 | 0.0037 | 68.3 | 0.1124 | 17.6 | 3.88 | 0.94 |
| BXH9_3_2G | BXH9 | 0.0042 | 131.3 | 0.1220 | 50.7 | 2.59 | 0.84 |
| HXB1-2mg-1 | HXB1 | 0.0061 | 42.9 | 0.1572 | 19.1 | 2.24 | 0.80 |
| HXB1-2mg-2 | HXB1 | 0.0095 | 101.1 | 0.2200 | 164.7 | 0.61 | 0.35 |
| HXB1-2mg-3 | HXB1 | 0.0098 | 105.4 | 0.1761 | 639.3 | 0.16 | 0.03 |
| HBX10_2_2G | HXB10 | 0.0032 | 16.3 | 0.0897 | 17.0 | 0.96 | 0.48 |
| HXB10_1_2G | HXB10 | 0.0047 | 64.9 | 0.1071 | 31.9 | 2.03 | 0.77 |
| HXB10_3_2G | HXB10 | 0.0036 | 36.7 | 0.0912 | 37.6 | 0.97 | 0.49 |
| HXB13_1_2G | HXB13 | 0.0011 | 83.5 | 0.0479 | 84.8 | 0.98 | 0.49 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| HXB13_3_2G | HXB13 | 0.0007 | 47.8 | 0.0336 | 43.9 | 1.09 | 0.54 |
| HXB15-2mg-1 | HXB15 | 0.0041 | 149.2 | 0.1215 | 182.2 | 0.82 | 0.41 |
| HXB15-2mg-2 | HXB15 | 0.0050 | 276.9 | 0.1143 | 159.2 | 1.74 | 0.74 |
| HXB15-2mg-3 | HXB15 | 0.0043 | 69.2 | 0.1018 | 28.7 | 2.41 | 0.85 |
| HXB17-2mg-1 | HXB17 | 0.0070 | 16.8 | 0.1130 | 94.5 | 0.18 | 0.06 |
| HXB17-2mg-2 | HXB17 | 0.0069 | 63.0 | 0.1869 | 33.9 | 1.86 | 0.72 |
| HXB17-2mg-3 | HXB17 | 0.0052 | 69.8 | 0.1436 | 10.8 | 6.48 | 0.97 |
| HXB18-2mg-1 | HXB18 | 0.0053 | 209.8 | 0.1614 | 98.7 | 2.13 | 0.76 |
| HXB18-2mg-2 | HXB18 | 0.0060 | 99.8 | 0.1390 | 25.9 | 3.86 | 0.92 |
| HXB18-2mg-3 | HXB18 | 0.0062 | 34.4 | 0.1184 | 10.5 | 3.27 | 0.86 |
| HXB2-2mg-1 | HXB2 | 0.0038 | 75.1 | 0.1044 | 104.1 | 0.72 | 0.35 |
| HXB2-2mg-2 | HXB2 | 0.0036 | 120.0 | 0.1192 | 47.1 | 2.54 | 0.86 |
| HXB2-2mg-3 | HXB2 | 0.0026 | 485.3 | 0.1039 | 423.1 | 1.15 | 0.57 |
| HXB20-2mg-1 | HXB20 | 0.0038 | 80.8 | 0.1107 | 63.4 | 1.27 | 0.61 |
| HXB20-2mg-2 | HXB20 | 0.0047 | 71.4 | 0.1110 | 21.2 | 3.37 | 0.92 |
| HXB20-2mg-3 | HXB20 | 0.0055 | 215.3 | 0.1271 | 134.8 | 1.60 | 0.67 |
| HXB21-2mg-1 | HXB21 | 0.0045 | 119.2 | 0.1491 | 54.0 | 2.21 | 0.82 |
| HXB21-2mg-2 | HXB21 | 0.0034 | 150.2 | 0.0951 | 77.6 | 1.94 | 0.76 |
| HXB21-2mg-3 | HXB21 | 0.0055 | 85.2 | 0.1393 | 28.4 | 3.00 | 0.84 |
| HXB22-2mg-1 | HXB22 | 0.0031 | 179.7 | 0.0823 | 224.2 | 0.80 | 0.40 |
| HXB22-2mg-2 | HXB22 | 0.0062 | 84.5 | 0.1371 | 168.7 | 0.50 | 0.21 |
| HXB22-2mg-3 | HXB22 | 0.0048 | 147.8 | 0.1229 | 50.3 | 2.94 | 0.89 |
| HXB23_1_2G | HXB23 | 0.0047 | 29.6 | 0.0906 | 11.2 | 2.65 | 0.82 |
| HXB23_2_2G | HXB23 | 0.0034 | 41.0 | 0.0799 | 33.7 | 1.22 | 0.57 |
| HXB23_3_2G | HXB23 | 0.0032 | 5.6 | 0.1067 | 1.3 | 4.25 | 0.87 |
| HXB24-2mg-1 | HXB24 | 0.0022 | 65.6 | 0.0718 | 52.0 | 1.26 | 0.60 |
| HXB24-2mg-2 | HXB24 | 0.0027 | 55.1 | 0.0967 | 38.5 | 1.43 | 0.65 |
| HXB24-2mg-3 | HXB24 | 0.0029 | 132.5 | 0.0890 | 94.2 | 1.41 | 0.66 |
| HXB25_1_2G | HXB25 | 0.0037 | 172.6 | 0.1009 | 112.3 | 1.54 | 0.69 |
| HXB25_3_2G | HXB25 | 0.0046 | 36.1 | 0.1222 | 94.4 | 0.38 | 0.19 |
| HXB27_1_2G | HXB27 | 0.0068 | 75.6 | 0.1598 | 7.9 | 9.51 | 0.97 |
| HXB27_2_2G | HXB27 | 0.0061 | 11.7 | 0.1676 | 16.5 | 0.71 | 0.39 |
| HXB27_3_2G | HXB27 | 0.0061 | 14.4 | 0.1745 | 36.3 | 0.40 | 0.23 |
| HXB29-2mg-1 | HXB29 | 0.0025 | 124.4 | 0.0745 | 148.1 | 0.84 | 0.42 |
| HXB29-2mg-2 | HXB29 | 0.0023 | 20.0 | 0.0614 | 16.8 | 1.19 | 0.57 |
| HXB29-2mg-3 | HXB29 | 0.0039 | 127.9 | 0.1128 | 148.0 | 0.86 | 0.43 |
| HXB3_1_2G | HXB3 | 0.0073 | 64.4 | 0.1653 | 7.3 | 8.80 | 0.97 |
| HXB3_2_2G | HXB3 | 0.0072 | 111.0 | 0.1423 | 52.6 | 2.11 | 0.78 |
| HXB31_1_2G | HXB31 | 0.0044 | 69.8 | 0.1016 | 37.5 | 1.86 | 0.77 |
| HXB31_3_2G | HXB31 | 0.0054 | 221.1 | 0.1449 | 322.4 | 0.69 | 0.32 |
| HXB4-2mg-1 | HXB4 | 0.0051 | 104.3 | 0.1423 | 35.3 | 2.95 | 0.89 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| HXB4-2mg-2 | HXB4 | 0.0039 | 145.1 | 0.1118 | 175.0 | 0.83 | 0.42 |
| HXB4-2mg-3 | HXB4 | 0.0050 | 146.8 | 0.1580 | 104.7 | 1.40 | 0.65 |
| HXB5-2mg-1 | HXB5 | 0.0050 | 82.7 | 0.1187 | 29.5 | 2.81 | 0.86 |
| HXB5-2mg-2 | HXB5 | 0.0063 | 70.0 | 0.1455 | 221.1 | 0.32 | 0.09 |
| HXB5-2mg-3 | HXB5 | 0.0063 | 39.6 | 0.1145 | 83.0 | 0.48 | 0.19 |
| HXB7_2_2G | HXB7 | 0.0056 | 11.2 | 0.0999 | 15.1 | 0.74 | 0.39 |
| HXB7_3_2G | HXB7 | 0.0029 | 109.5 | 0.0845 | 64.7 | 1.69 | 0.73 |

Each row represents an individual animal. For each animal, the elimination phase of ethanol, i.e., the concentration-time curve from peak blood concentration through the remaining data after a 2 g/kg dose of ethanol, was fit to a zero-order (i.e., straight line) and first-order (i.e., exponential decay) kinetic model and the corresponding rate constant calculated. The F-statistic was determined by taking the ratio of the sum of residuals for the first-order fit to the zero-order fit for each individual animal. A small p-value for the F-statistic indicates the first-order model fit the concentration-time curve more accurately for the given animal, and a large p-value indicates the zero-order model was a better fit. The vast majority of p-values for the rats lie between 0.05 and 0.95, which suggests neither model fit the animal significantly better than the other.

## Supplementary Table S2. Detailed summary of RNA sequencing results by sample

| Batch 1 | | | | | Batch 2 | | | | | Batch 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | Number of Paired-End Reads | Number of Paired-End Reads After Trimming | Percent of Paired-End Reads Aligned to rRNA | Number of Paired-End Reads NOT Aligned to rRNA | Sample | Number of Paired-End Reads | Number of Paired-End Reads After Trimming | Percent of Paired-End Reads Aligned to rRNA | Number of Paired-End Reads NOT Aligned to rRNA | Sample | Number of Paired-End Reads | Number of Paired-End Reads After Trimming | Percent of Paired-End Reads Aligned to rRNA | Number of Paired-End Reads NOT Aligned to rRNA |
| BXH12_1 | 48944095 | 46522887 | 1.1% | 45993223 | BXH2 | 42903025 | 42775273 | 0.6% | 42536334 | BNLx_1 | 90666458 | 90074574 | 1.0% | 89129996 |
| BXH12_2 | 46809812 | 44375871 | 3.4% | 42854944 | HXB10 | 47121630 | 46941747 | 0.7% | 46634733 | BNLx_2 | 59006131 | 58598172 | 0.8% | 58106969 |
| HXB13_1 | 50150896 | 46695549 | 1.3% | 46073379 | HXB1 | 42944312 | 42818522 | 1.1% | 42349636 | BXH10_1 | 86114764 | 85143297 | 1.1% | 84243361 |
| HXB13_2 | 45529860 | 43140610 | 5.8% | 40647600 | HXB15 | 53558009 | 53443809 | 0.7% | 53084566 | BXH10_2 | 70514093 | 70140982 | 1.2% | 69302759 |
| HXB17_1 | 46594113 | 43848524 | 1.2% | 43332445 | HXB18 | 38623372 | 38468664 | 1.1% | 38040549 | BXH11_1 | 84084666 | 83482541 | 0.8% | 82832449 |
| HXB17_2 | 37696059 | 34782463 | 1.2% | 34374408 | HXB20 | 53698856 | 53408926 | 0.8% | 52972074 | BXH11_2 | 96941474 | 96387388 | 0.6% | 95796252 |
| HXB2_1 | 40389886 | 38408290 | 8.2% | 35247605 | HXB21 | 46122637 | 45786609 | 0.6% | 45512955 | BXH13_1 | 76910934 | 76398952 | 1.3% | 75423998 |
| HXB2_2 | 41680207 | 38891718 | 1.0% | 38492580 | HXB22 | 38836500 | 38660060 | 0.7% | 38389294 | BXH3_1 | 89250975 | 88220526 | 1.6% | 86851878 |
| HXB25_1 | 35104202 | 33348011 | 1.9% | 32708490 | HXB23 | 54376269 | 54248472 | 0.6% | 53922836 | BXH3_2 | 82735428 | 82205766 | 0.7% | 81646781 |
| HXB25_2 | 47102360 | 45162710 | 11.7% | 39859957 | HXB24 | 48470543 | 48325150 | 0.7% | 47999114 | BXH5_1 | 79983954 | 79290372 | 0.8% | 78693651 |
| HXB27_1 | 35651237 | 33631341 | 1.2% | 33212524 | HXB29 | 44560164 | 44415617 | 0.7% | 44099219 | BXH6_1 | 87032370 | 86490639 | 2.3% | 84540498 |
| HXB27_2 | 45626177 | 43339409 | 6.0% | 40737697 | HXB31 | 47936946 | 47786908 | 0.6% | 47489523 | BXH6_2 | 85302331 | 84720856 | 0.8% | 84017266 |
| HXB7_1 | 53174270 | 50102243 | 8.9% | 45637827 | HXB3 | 40526076 | 40398533 | 1.2% | 39931495 | BXH8_1 | 72771217 | 72054807 | 1.1% | 71292055 |
| HXB7_2 | 49021976 | 46849723 | 1.0% | 46376933 | HXB4 | 43726307 | 43612129 | 1.0% | 43169045 | BXH8_2 | 73575935 | 72444118 | 0.7% | 71929130 |
| SHR_1 | 49060694 | 47297561 | 12% | 41635176 | HXB5 | 49348173 | 49214553 | 1.0% | 48705074 | BXH9_1 | 102849405 | 102221494 | 1.3% | 100925554 |
| SHR_2 | 50009902 | 47699506 | 2.9% | 46336265 | SHR | 47027334 | 46839830 | 7.8% | 43167544 | BXH9_2 | 71741565 | 71367904 | 1.0% | 70645598 |
| | | | | | | | | | | HXB10_1 | 89785465 | 89027890 | 0.9% | 88256121 |
| | | | | | | | | | | HXB18_1 | 89184507 | 88827405 | 1.8% | 87251073 |
| | | | | | | | | | | HXB22_1 | 81854638 | 81416829 | 1.4% | 80263351 |
| | | | | | | | | | | HXB23_1 | 67289160 | 66497693 | 0.9% | 65926737 |
| | | | | | | | | | | HXB31_1 | 73540759 | 73100949 | 1.9% | 71743480 |
| | | | | | | | | | | HXB3_1 | 81953604 | 81299891 | 1.3% | 80220536 |
| | | | | | | | | | | SHR_1 | 70590530 | 70000608 | 0.9% | 69403209 |
| | | | | | | | | | | SHR_2 | 63996267 | 63304220 | 0.7% | 62837673 |

Samples are denoted by strain and if required sample number separated by an underscore.

**Supplementary Table S3A. Modules from HXB/BXH WGCNA that are associated with alcohol clearance in the recombinant inbred rat panel after 2 g/kg alcohol administration.**

| Module | Number of Genes in Module | Proportion of Variance in Module Explained By Eigengene | Correlation with Alcohol Clearance in HXB/BXH | | Maximum Module Eigengene QTL | |
|---|---|---|---|---|---|---|
| | | | Correlation Coefficient | p-value | Location [Chromosome:Mb (95% Bayesian Credible Interval)] | Empirical Genome-wide p-value |
| orange3 | 10 | 0.60 | -0.75 | <0.001 | chr2:242.9 (242.9-246.2) | < 0.001 |
| darkslateblue.1 | 6 | 0.66 | -0.56 | 0.0016 | chr3:36.4 (35.4-41.2) | 0.014 |
| peachpuff3 | 9 | 0.59 | -0.55 | 0.0018 | chr5:54.7 (1.8-162.6) | 0.97 |
| palegreen1 | 8 | 0.62 | -0.53 | 0.0030 | chr6:54.3 (0.1-121.5) | 0.63 |
| maroon4 | 8 | 0.63 | 0.50 | 0.0055 | chr2:242.9 (2.6-265.4) | 0.13 |
| lightcyan.1 | 6 | 0.65 | -0.49 | 0.0065 | chr7:108.9 (11.8-144.4) | 0.89 |
| mistyrose3.1 | 5 | 0.66 | 0.49 | 0.0075 | chr6:110.1 (24.8-145.1) | 0.54 |
| palegreen3 | 8 | 0.66 | 0.49 | 0.0076 | chr3:41.2 (35.4-58.1) | 0.032 |
| darkorange3 | 7 | 0.68 | 0.48 | 0.0084 | chr12:33.1 (2.2-47.8) | 0.51 |
| goldenrod2 | 8 | 0.62 | -0.48 | 0.0091 | chr3:45.3 (4.1-176.6) | 0.60 |
| | | | | | | |
| | | | | | | |

Module eigengenes were used to determine correlation (Pearson) between modules and alcohol clearance. Modules with a correlation coefficient p-value < 0.01 were included and are ordered by p-value. The location and empirical genome-wide p-value of the most significant module eigengene QTL for each module is included.

**Supplementary Table S3B. Modules from HXB/BXH WGCNA that are associated with acetate area under the curve (AUC) in the recombinant inbred rat panel after 2 g/kg alcohol administration.**

| Module | Number of Genes in Module | Proportion of Variance in Module Explained By Eigengene | Correlation with Acetate AUC in HXB/BXH | | Maximum Module Eigengene QTL | |
|---|---|---|---|---|---|---|
| | | | Correlation Coefficient | p-value | Location [Chromosome:Mb (95% Bayesian Credible Interval)] | Empirical Genome-wide p-value |
| lightslateblue | 14 | 0.54 | 0.70 | <0.001 | chr18:54.6 (54.1-56.4) | 0.013 |
| sienna2.1 | 5 | 0.68 | 0.62 | 0.0003 | chr2:229.3 (2.6-243.8) | 0.20 |
| bisque3.1 | 5 | 0.65 | -0.58 | 0.0009 | chr19:49.5 (10.9-60.3) | 0.49 |
| orange3 | 10 | 0.60 | -0.57 | 0.0014 | chr2:242.9 (242.9-246.2) | < 0.001 |
| antiquewhite4 | 22 | 0.55 | 0.51 | 0.0050 | chr2:186.3 (186.3-186.3) | < 0.001 |
| goldenrod2 | 8 | 0.62 | -0.51 | 0.0051 | chr3:45.3 (4.1-176.6) | 0.60 |
| cadetblue4 | 7 | 0.70 | 0.50 | 0.0063 | chr3:40.5 (33.1-108.6) | 0.14 |
| maroon4 | 8 | 0.63 | 0.49 | 0.0073 | chr2:242.9 (2.6-265.4) | 0.13 |
| deepskyblue2 | 8 | 0.64 | 0.47 | 0.0095 | chr13:45.3 (18.1-72.5) | 0.065 |
| cadetblue | 8 | 0.62 | -0.47 | 0.0100 | chr18:6.8 (0.4-15.3) | 0.22 |

Module eigengenes were used to determine correlation (Pearson) between modules and acetate AUC. Modules with a correlation coefficient p-value < 0.01 were included and are ordered by p-value. The location and empirical genome-wide p-value of the most significant module eigengene QTL for each module is included.

**Supplementary Table S4. Associations between Ensembl genes, Ensembl transcripts and the phenotypes for the alcohol dehydrogenase genes in the orange3 candidate module.**

| | *Adh 6* | ENSRNOT00000036993 (*Adh6*) | ENSRNOT00000008382 (*Adh1*) | ENSRNOT00000085067 (Fusion Gene with *Adh1* and *Adh6*) | *Adh 4* | ENSRNOT00000016891 (*Adh4*) | ENSRNOT00000017252 (*Adh5*) | ENSRNOT00000090253 (Fusion Gene with *Adh4* and *Adh5*) | Acetate AUC | Alcohol Clearance |
|---|---|---|---|---|---|---|---|---|---|---|
| *Adh6* | 1.00 | 0.86 | **0.95** | 0.21 | 0.89 | 0.85 | 0.43 | 0.59 | 0.57 | 0.76 |
| ENSRNOT00000036993 (*Adh6*) | 0.86 | 1.00 | 0.77 | 0.30 | 0.92 | 0.90 | 0.55 | 0.69 | 0.53 | 0.64 |
| ENSRNOT00000083682 (*Adh1*) | 0.95 | 0.77 | 1.00 | 0.23 | 0.80 | 0.74 | 0.52 | 0.42 | 0.52 | 0.72 |
| ENSRNOT00000085067 (Fusion Gene with *Adh1* and *Adh6*) | 0.21 | 0.30 | 0.23 | 1.00 | 0.33 | 0.33 | 0.06 | 0.02 | 0.35 | 0.06 |
| *Adh4* | 0.89 | 0.92 | 0.80 | 0.33 | 1.00 | **0.97** | 0.48 | 0.67 | 0.52 | 0.64 |
| ENSRNOT00000016891 (*Adh4*) | 0.85 | 0.90 | 0.74 | 0.33 | 0.97 | 1.00 | 0.33 | 0.66 | 0.46 | 0.63 |
| ENSRNOT00000017252 (*Adh5*) | 0.43 | 0.55 | 0.52 | 0.06 | 0.48 | 0.33 | 1.00 | 0.33 | 0.20 | 0.23 |
| ENSRNOT00000090253 (Fusion Gene with *Adh4* and *Adh5*) | 0.59 | 0.69 | 0.42 | 0.02 | 0.67 | 0.66 | 0.33 | 1.00 | 0.47 | 0.41 |
| Acetate AUC | **0.57** | 0.53 | **0.52** | 0.35 | **0.52** | **0.46** | 0.20 | 0.47 | 1.00 | 0.44 |
| Alcohol Clearance | **0.76** | 0.64 | **0.72** | 0.06 | **0.64** | **0.63** | 0.23 | 0.41 | 0.44 | 1.00 |

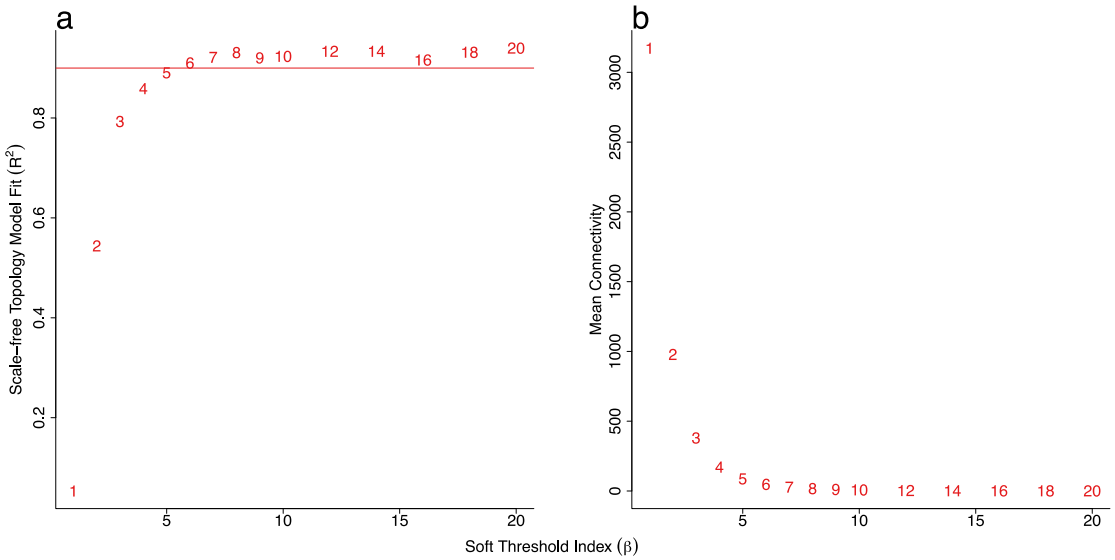
The Ensembl genes identified in the orange3 module are listed by themselves (Adh6 = alcohol dehydrogenase 6 Adh4 = alcohol dehydrogenase 4) and the three Ensembl transcripts whose pooled expression yielded the overall Ensembl gene expression estimate (ENSRNOT00000036993 ENSRNOT00000083682 and ENSRNOT00000085067 for Adh6 and ENSRNOT00000016891 ENSRNOT00000017252 and ENSRNOT00000090253 for Adh4) are listed with the corresponding RefSeq gene annotations in parenthesis (those listed as fusion genes did not have a corresponding RefSeq gene annotation). Cells with bold text indicate 1) the similarity in expression between the overall Ensembl gene and a single Ensembl transcript and 2) the similarity in the association of expression with the phenotype(s) between the overall Ensembl gene and a single Ensembl transcript across the strains of the HXB/BXH recombinant inbred rat panels. Strain means were used for pairwise Pearson correlation analysis.

**Supplementary Table S5. Comparison of expression and expression associations with alcohol clearance and acetate area under the curve (AUC) across the HXB/BXH recombinant inbred (RI) rat panel for selected genes.**

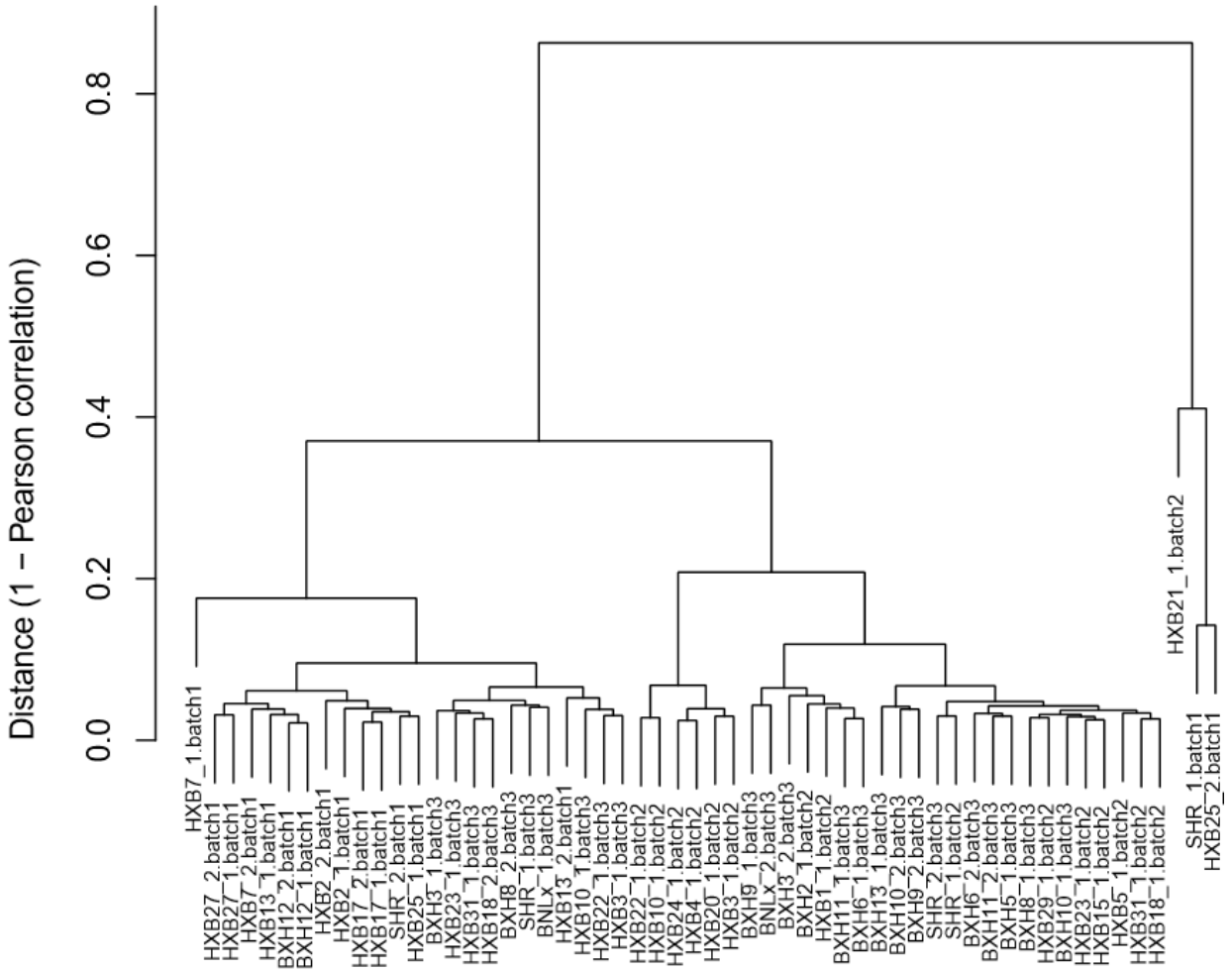
| Gene | Total Counts in HXB/BXH | Ratio of Expression in HXB/BXH | Correlation of Expression and Alcohol Clearance in HXB/BXH | | Correlation of Expression and Acetate AUC in HXB/BXH | | Module Membership | Description |
|---|---|---|---|---|---|---|---|---|
| | | | Correlation Coefficient | p-value | Correlation Coefficient | p-value | | |
| *Adh1* | 716112 | 3.5 | 0.76 | **<0.001** | 0.57 | **0.0012** | orange3 | Alcohol dehydrogenase 1 (class I) |
| *Adh4* | 180249 | 2.9 | 0.64 | **<0.001** | 0.52 | **0.0037** | orange3 | Alcohol dehydrogenase 4 (class II) pi polypeptide |
| *Adh7* | 10056 | 2.3 | 0.28 | 0.14 | 0.43 | 0.019 | green | Alcohol dehydrogenase 7 (class IV) mu or sigma polypeptide |
| *Adhfe1* | 75819 | 3.1 | -0.07 | 0.72 | -0.04 | 0.82 | pink3.1 | Alcohol dehydrogenase iron containing 1 |
| *Cyp2e1* | 4428563 | 2.0 | 0.21 | 0.26 | 0.02 | 0.92 | plum2 | Cytochrome P450 family 2 subfamily e polypeptide 1 |
| *Cat* | 1111940 | 2.0 | -0.27 | 0.16 | 0.11 | 0.55 | turquoise | Catalase |
| *Aldh2* | 64713 | 2.6 | -0.20 | 0.30 | 0.09 | 0.64 | turquoise | Aldehyde dehydrogenase 2 family (mitochondrial) |
| *Aldh3a2* | 445668 | 1.5 | 0.06 | 0.75 | -0.21 | 0.27 | steelblue | Aldehyde dehydrogenase 3 family member A2 |
| *Aldh9a1* | 266873 | 1.5 | 0.00 | 1.00 | -0.17 | 0.37 | mediumorchid3 | Aldehyde dehydrogenase 9 family member A1 |
| *Aldh1l2* | 23678 | 2.8 | -0.19 | 0.31 | -0.02 | 0.93 | yellow | Aldehyde dehydrogenase 1 family member L2 |
| *Aldh6a1* | 733434 | 1.4 | -0.04 | 0.84 | 0.11 | 0.57 | lavenderblush2 | Aldehyde dehydrogenase 6 family member A1 |
| *Aldh1b1* | 18735 | 2.0 | 0.13 | 0.52 | 0.03 | 0.87 | brown | Aldehyde dehydrogenase 1 family member B1 |
| *Aldh7a1* | 248254 | 1.5 | -0.03 | 0.88 | -0.07 | 0.71 | tan2 | Aldehyde dehydrogenase 7 family member A1 |
| *Aldh8a1* | 199173 | 1.3 | -0.38 | 0.044 | 0.00 | 0.98 | orange | Aldehyde dehydrogenase 8 family member A1 |
| *Aldh18a1* | 162 | 2.1 | 0.00 | 0.99 | -0.16 | 0.40 | darkolivegreen3 | Aldehyde dehydrogenase 18 family member A1 |
| *Aldh3b1* | 466 | 1.4 | 0.02 | 0.93 | -0.05 | 0.80 | brown | Aldehyde dehydrogenase 3 family member B1 |
| *Aldh1a1* | 331535 | 3.1 | 0.14 | 0.48 | 0.36 | 0.052 | violet | Aldehyde dehydrogenase 1 family member A1 |
| *Aldh1a7* | 19223 | 42.5 | -0.12 | 0.53 | 0.15 | 0.45 | brown.1 | Aldehyde dehydrogenase family 1 subfamily A7 |
| *Aldh16a1* | 4985 | 1.4 | -0.16 | 0.39 | 0.01 | 0.96 | cornsilk2 | Aldehyde dehydrogenase 16 family member A1 |
| *Aldh5a1* | 60984 | 1.4 | 0.22 | 0.25 | 0.15 | 0.44 | blue | Aldehyde dehydrogenase 5 family member A1 |
| *Aldh1l1* | 104751 | 2.8 | 0.01 | 0.96 | 0.02 | 0.90 | turquoise | Aldehyde dehydrogenase 1 family member L1 |
| *Aldh1a3* | 18433 | 2.0 | -0.11 | 0.57 | -0.14 | 0.47 | lightgoldenrod1 | Aldehyde dehydrogenase 1 family member A3 |
| *Aldh1a2* | 580 | 1.4 | 0.15 | 0.43 | -0.07 | 0.71 | turquoise | Retinal dehydrogenase 2 |

Total counts in HXB/BXH for each gene represent the summed counts across all liver samples used to derive final strain mean expression estimates (41 samples total) using the expression values after normalization via removal of unwanted variance but prior to rlog-transformation and strain averaging. For the ratio of expression in the HXB/BXH RI rat panel the final strain mean rlog-transformed expression estimates used to build co-expression networks were back transformed and the highest expression estimate in any strain over the smallest expression estimate in any strain across the HXB/BXH RI panel was calculated for each gene. Pairwise Pearson correlation analysis on strain mean values of gene expression estimates alcohol clearance and acetate AUC was used to determine correlation coefficients. Significant (nominal p-value < 0.05) associations are denoted with bold text.

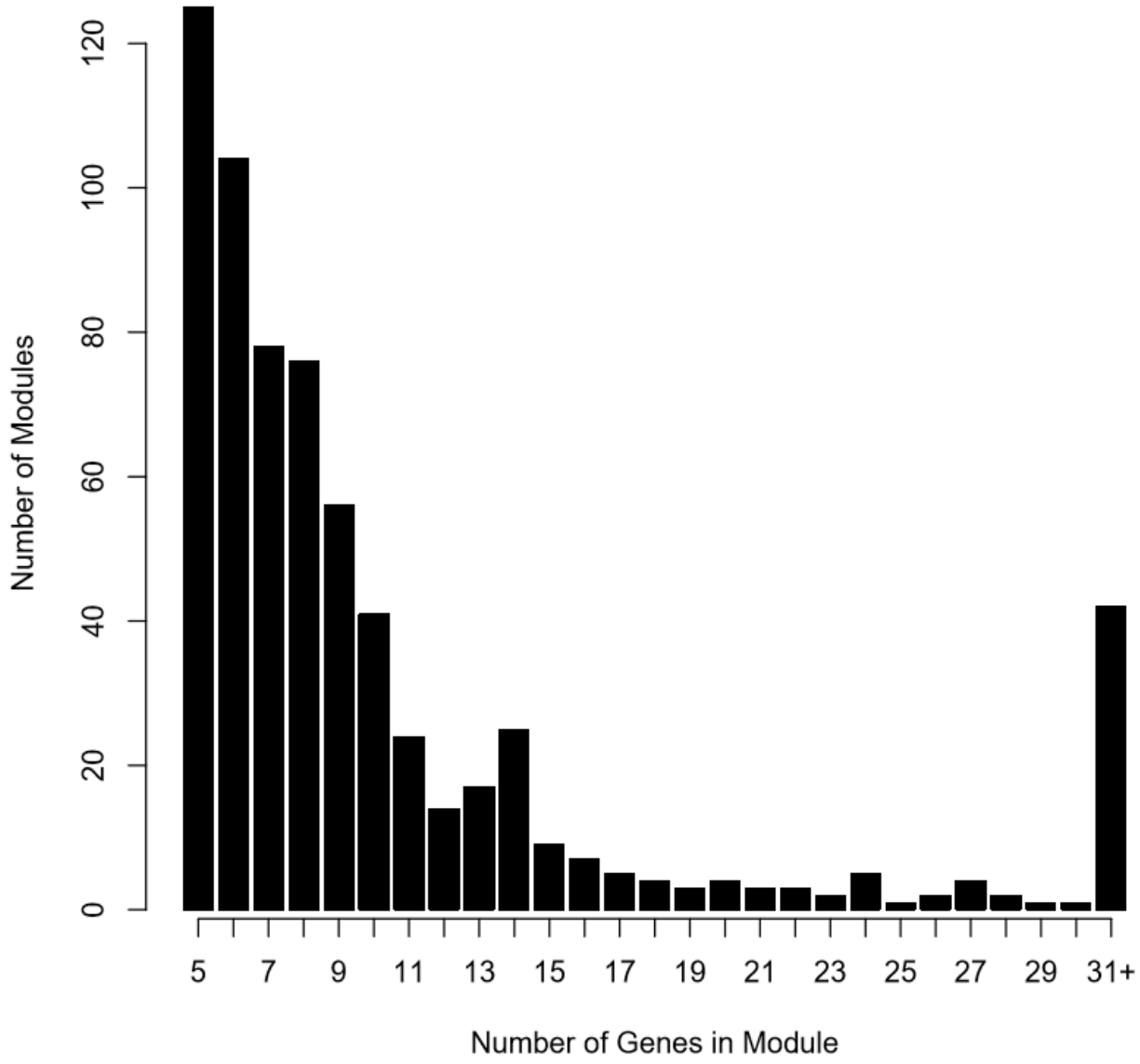
## Supplementary Figures



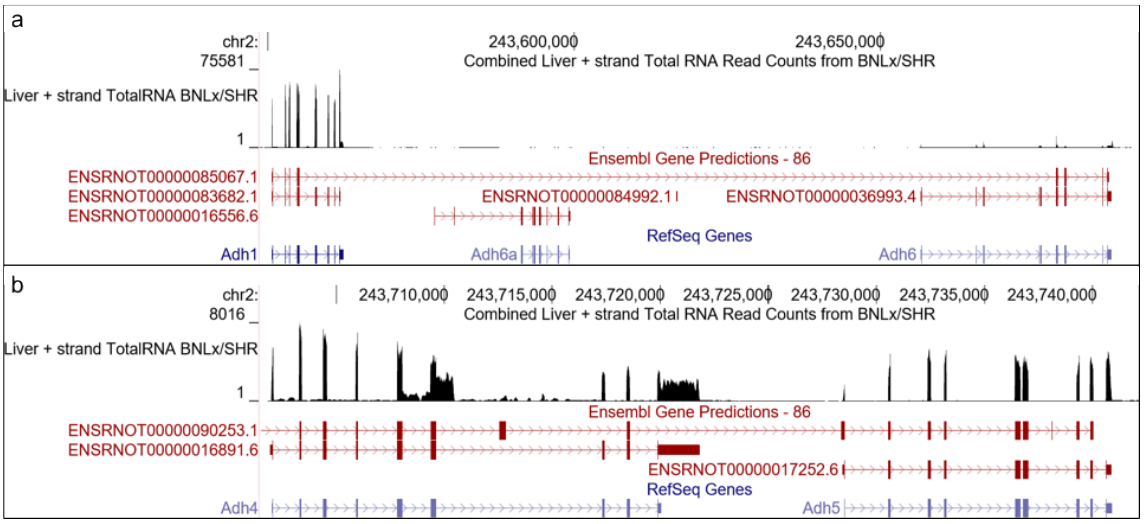

**Supplementary Figure S1. Network topology as a function of soft-thresholding power (β).** The influence of different index values on the goodness-of-fit to scale-free topology and topological features for the network were examined in **(a) scale-free model fit index vs β and (b) mean connectivity vs β.** Each point is labeled by its β index value in red. In (a), the red line represents a recommended cutoff value of 0.90 (Zhang and Horvath 2005; Stat Appl Genet Mol Biol. 4:17). An index of six was chosen for network construction.



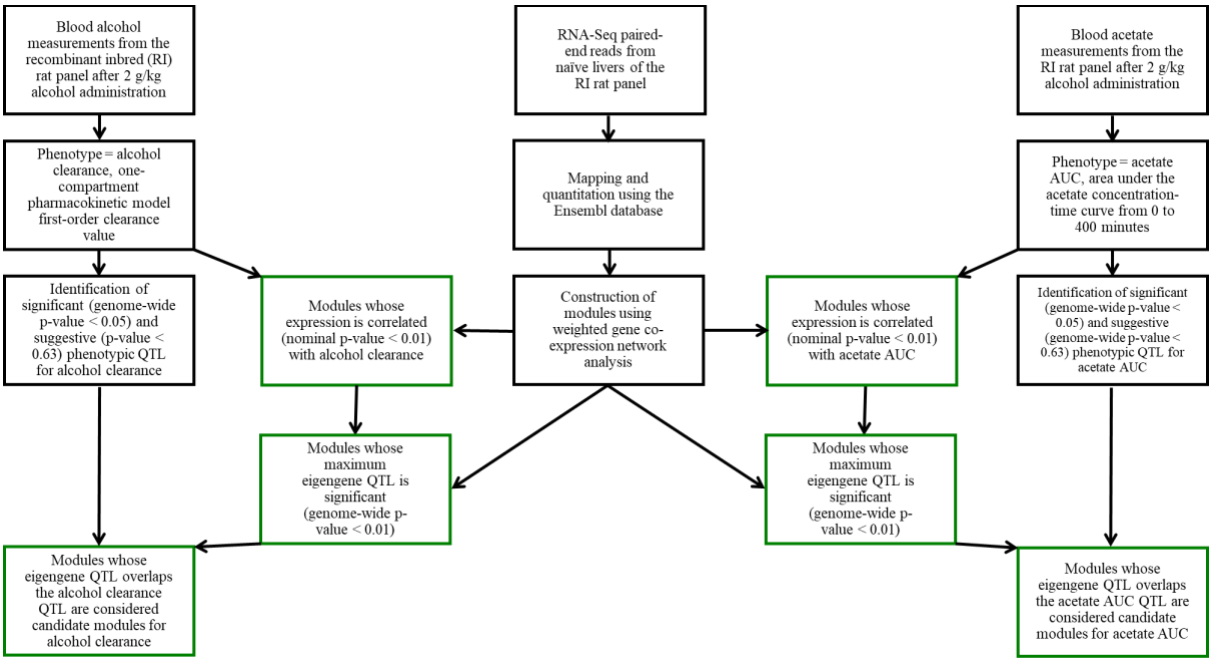

**Supplementary Figure S2. Clustering of liver samples after quantification of gene abundances.** Distances were calculated using Pearson correlation on the log2(RSEM counts + 1) transformed gene expression data. Leaves of the dendrogram are labeled by strain, sample number, and batch (strain_sample number.batch). Samples HXB21_1.batch2, SHR_1.batch1, HXB25_2.batch1, and HXB7_1.batch1 were identified as outliers and removed from subsequent analyses.



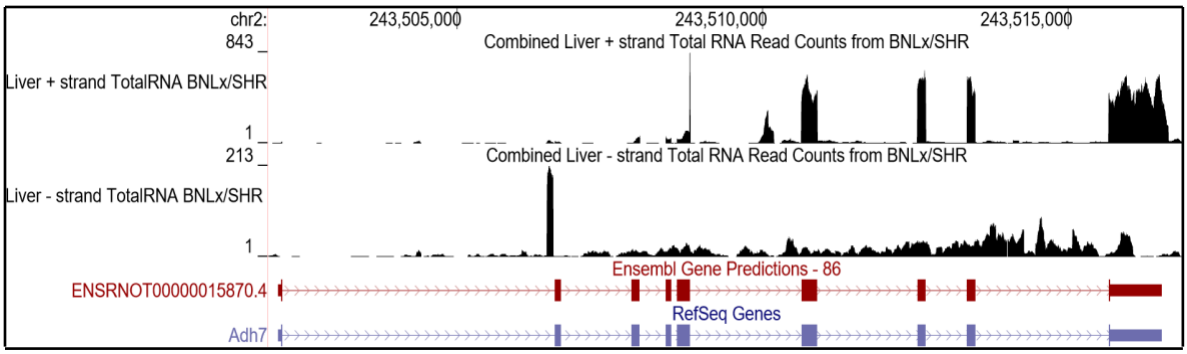

**Supplementary Figure S3. Distribution of module sizes built from weighted gene co-expression network analysis of the liver RNA expression data in the HXB/BXH recombinant inbred rat panel.** Module size is defined as the number of genes in the module. Module with more than 30 genes were put into a single bin (31+).




**Supplementary Figure S4. Alignment of the RNA sequencing (RNA-Seq) reads from the plus strand of the progenitor strains to the Rnor_6.0 rat genome locations containing Ensembl and/or RefSeq annotations for (a) alcohol dehydrogenase 1 (*Adh1*), alcohol dehydrogenase 6 (*Adh6*), and the fusion of these genes** and for **(b) alcohol dehydrogenase 4 (*Adh4*), alcohol dehydrogenase 5 (*Adh5*), and the fusion of these genes.** Ensembl and RefSeq annotations are labeled in red and blue, respectively, where blocks represent exons. The track representing the liver total RNA-Seq reads from the plus strand of the progenitor strains (BN-Lx/Cub and SHR/OlaIpcv) is labeled to the left in the figures with the corresponding pile up of reads represented in black. The negative strand was not significantly expressed and as a result not included. In (a), the Ensembl annotations ENSRNOT00000083682.1 (far left), ENSRNOT00000036993.4 (far right), and ENSRNOT00000085067.1 correspond to the three Ensembl transcripts whose pooled expression estimates produced the gene-level Ensembl *Adh6* estimate. Of these three transcripts, the vast majority of reads align to ENSRNOT00000083682.1, which in RefSeq is annotated as its own unique gene, *Adh1*. In (b), the Ensembl transcripts whose pooled expression estimates yielded the gene-level Ensembl *Adh4* estimate are ENSRNOT00000016891.6 (far left), ENSRNOT00000017252.6 (far right), and ENSRNOT00000090253.1. Here the distribution of reads is split between the two Ensembl transcripts ENSRNOT00000016891.6 and ENSRNOT00000017252.6 that are annotated in RefSeq as two unique genes *Adh4* and *Adh5*, respectively; however, correlation analysis between the Ensembl gene-level *Adh4*, Ensembl transcripts, and the phenotypes suggest Ensembl transcript/RefSeq ENSRNOT00000016891.6/*Adh4* is the main contributor to the variation in the overall Ensembl Adh4 gene-level expression. These images were generated using the UCSC genome browser (https://genome.ucsc.edu).



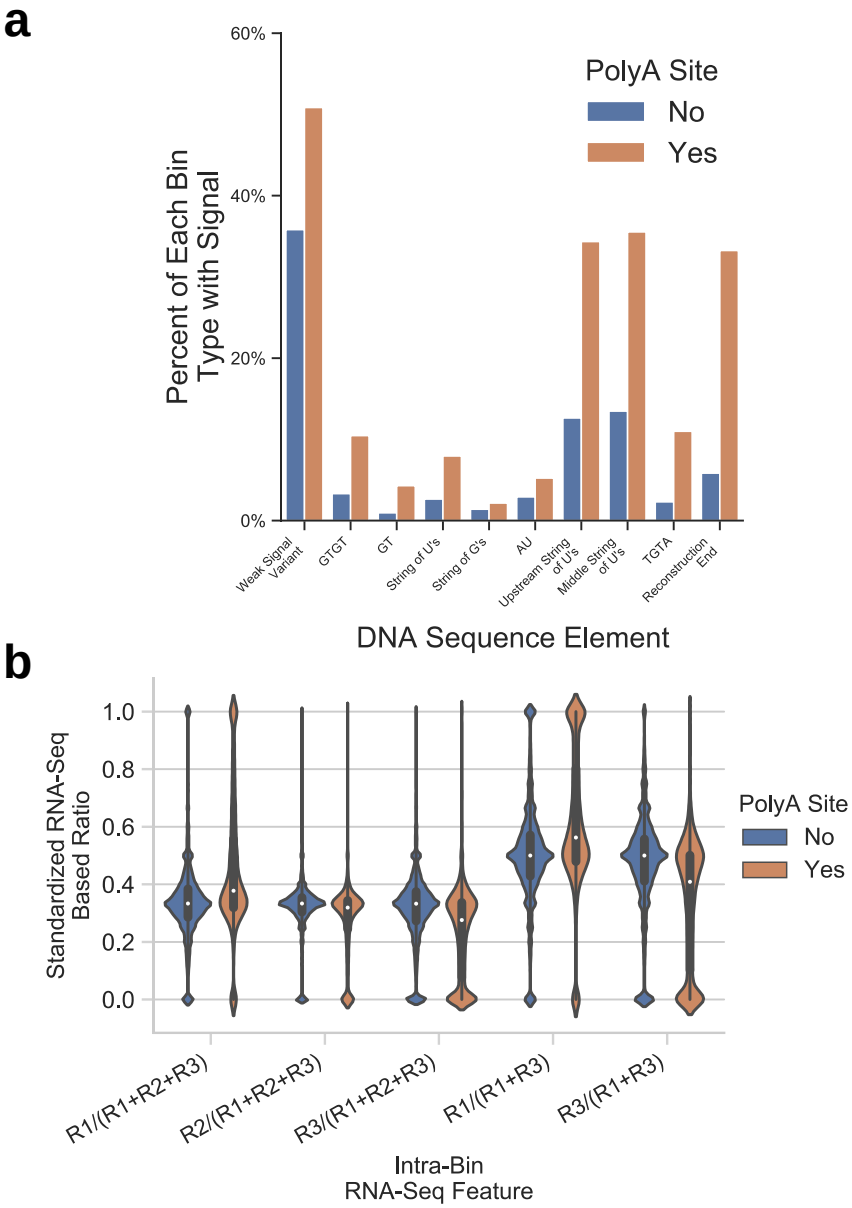

**Supplementary Figure S5. Workflow of methods used to identify candidate modules.** Boxes outlined in green denote steps where modules were filtered.



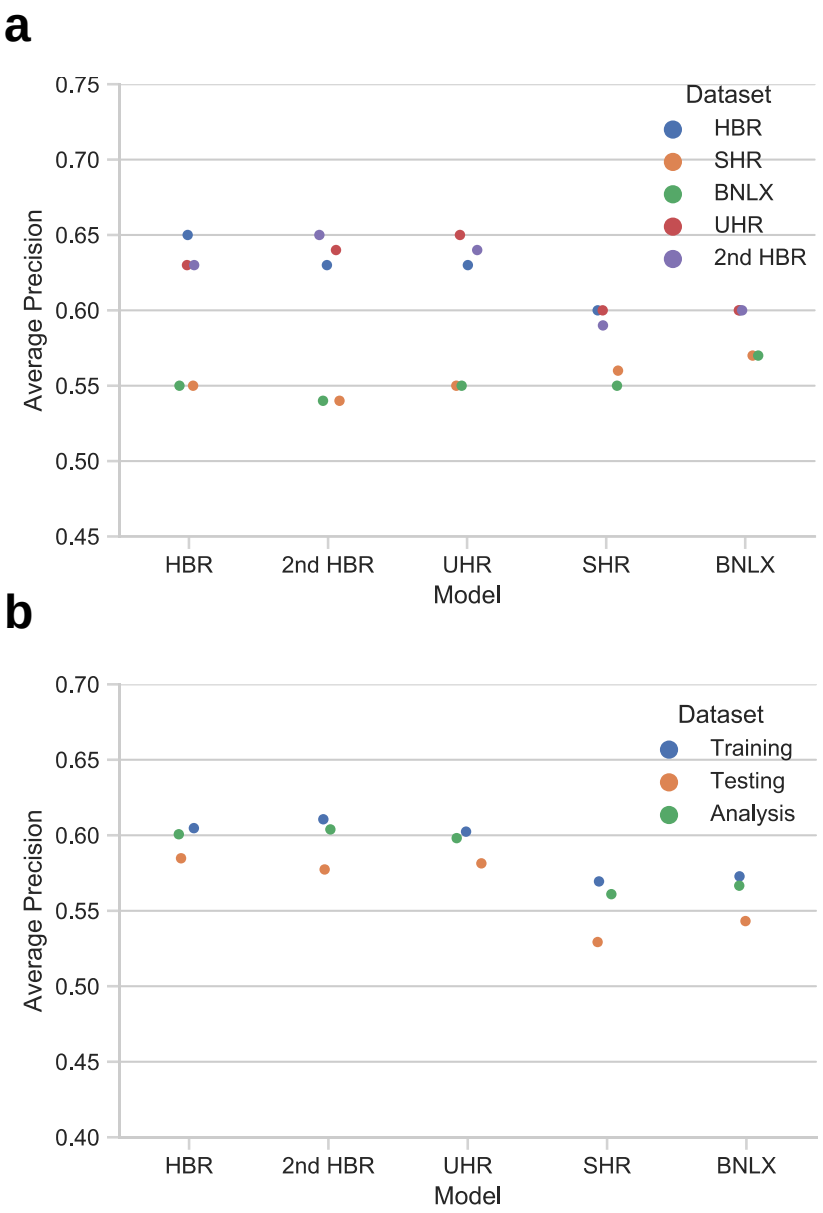

**Supplementary Figure S6. Alignment of the RNA sequencing (RNA-Seq) reads from the plus and minus strands of the progenitor strains to the Rnor_6.0 rat genome location containing the Ensembl and RefSeq annotations for alcohol dehydrogenase 7 (*Adh7*).** Ensembl and RefSeq annotations are labeled in red and blue, respectively, where blocks represent exons (ENSRNOT00000015870.4 is the single Ensembl transcript for *Adh7*). The tracks representing the liver total RNA-Seq reads from the plus and minus strands of the progenitor strains (BN-Lx/Cub and SHR/OlaIpcv) are labeled to the left in the figure with the corresponding pile up of reads represented in black. This image was generated using the UCSC genome browser (https://genome.ucsc.edu).

**APPENDIX B**

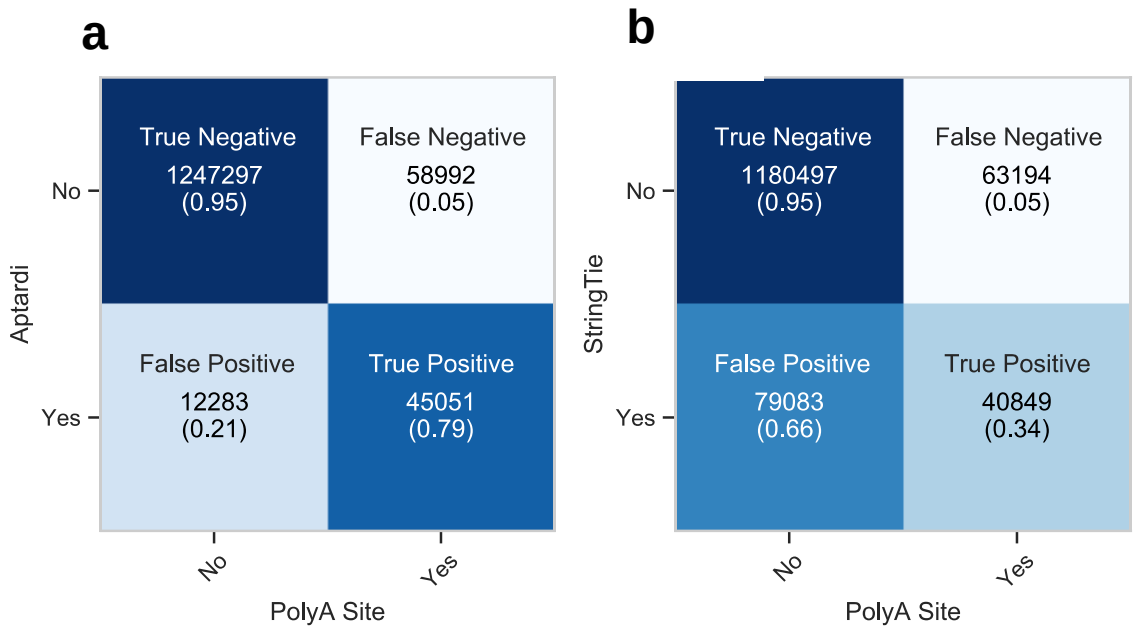
**CHAPTER III SUPPLEMENTARY**



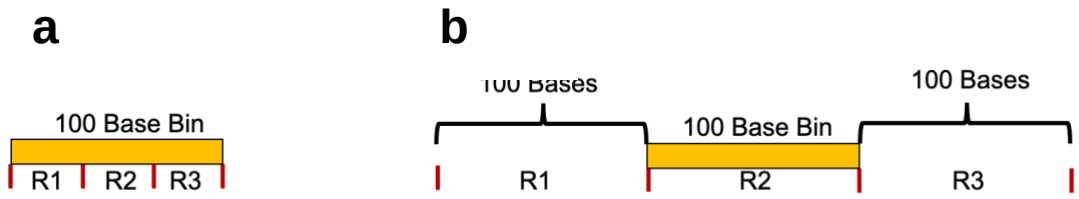

**Supplementary Fig. 1: DNA sequence and RNA sequencing (RNA-Seq) features for each bin as a function of whether the bin contains a polyadenylation (polyA) site. a,** The percent of 100 base bins containing the listed DNA sequence feature stratified by the bin not containing (blue) or containing (orange) a polyA site. **b,** Distribution of the standardized ratios for the intra-bin RNA-Seq features for each 100 base bin stratified by the bin not containing (blue) or containing (orange) a polyA site (each RNA-Seq ratio feature was standardized using the training set). Data shown are from the Human Brain Reference dataset.



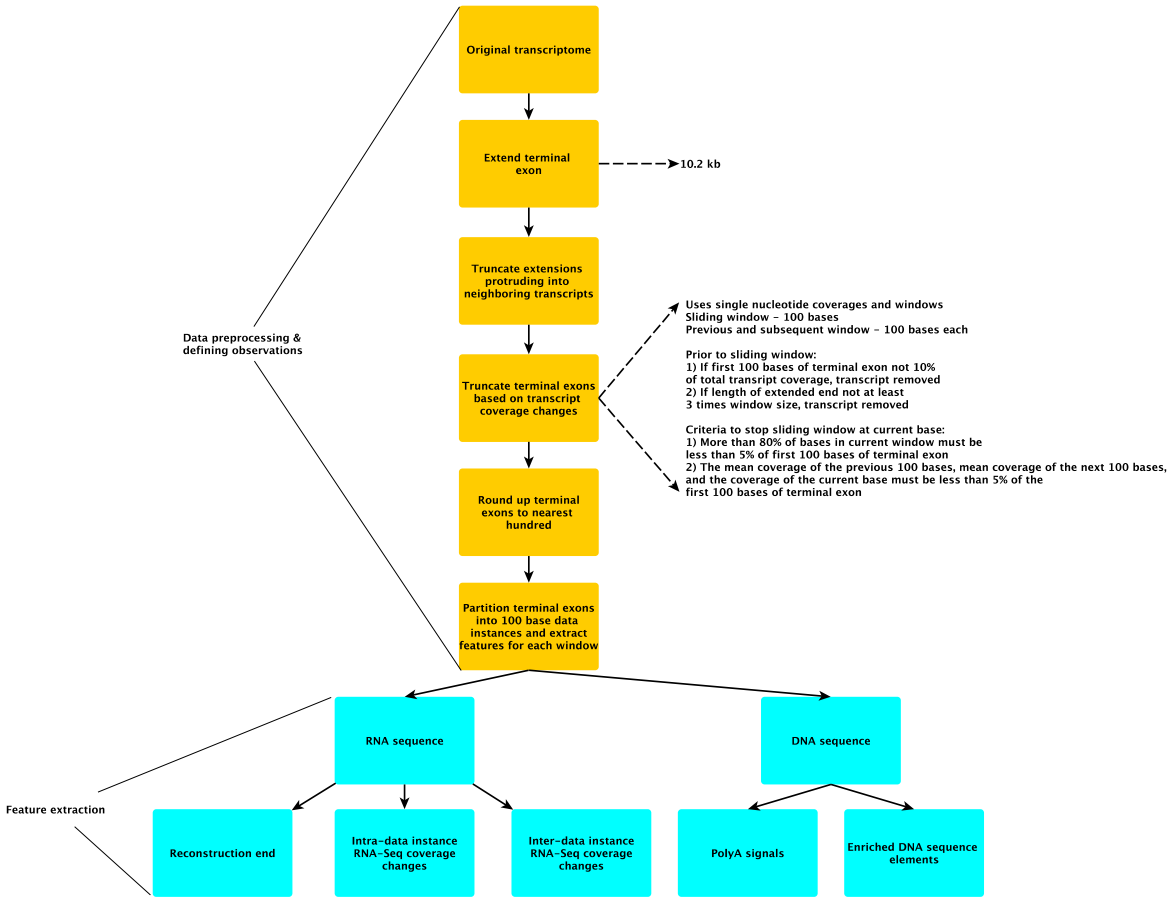

**Supplementary Fig. 2: The machine learning pipeline used to build aptardi is robust to different datasets. a**, Prediction models built on a given dataset perform comparably across all datasets. Colors denote the dataset used to build the predictive model, and the x-axis indicates the model used to calculate the average precision (y-axis) on the given dataset. **b**, Model performance is consistent similar the training, testing, and analysis (entire dataset without merging modified 3' terminal exons) sets.



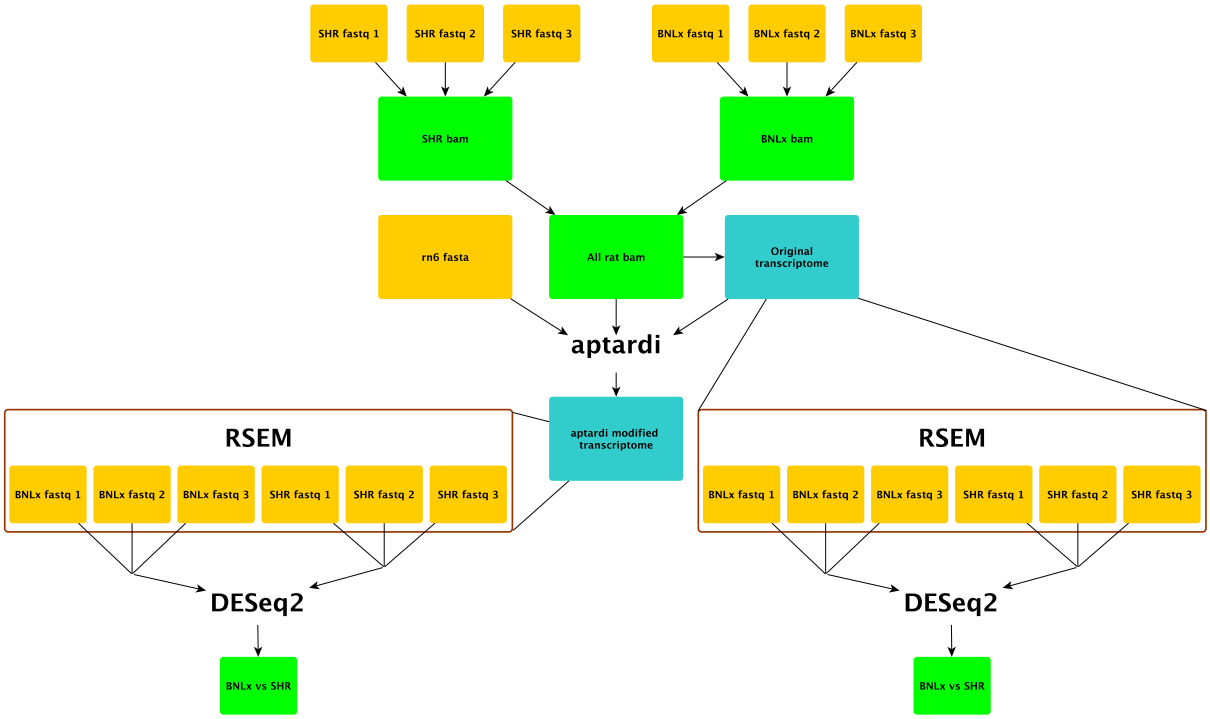

**Supplementary Fig. 3: Aptardi improves the classification confusion matrix compared to StringTie. a,** The confusion matrix from the aptardi prediction model generated from the Human Brain Reference (HBR) dataset improved the positive predictive value by increasing the proportion of true positive tests among positive aptardi results compared to **b,** the confusion matrix from StringTie on the same dataset. Classifications on each 100 base increment (i.e. bin) included in the analysis were compared. For the aptardi prediction model, its predictions for the presence (Yes) or absence (No) of a polyadenylation (polyA site) site were determined using the default probability threshold (0.5). For StringTie, the presence or absence of any 3' terminus within the bin from its transcriptome was used as positive and negative predictions, respectively. True polyA sites were taken from the HBR PolyA-Seq data.



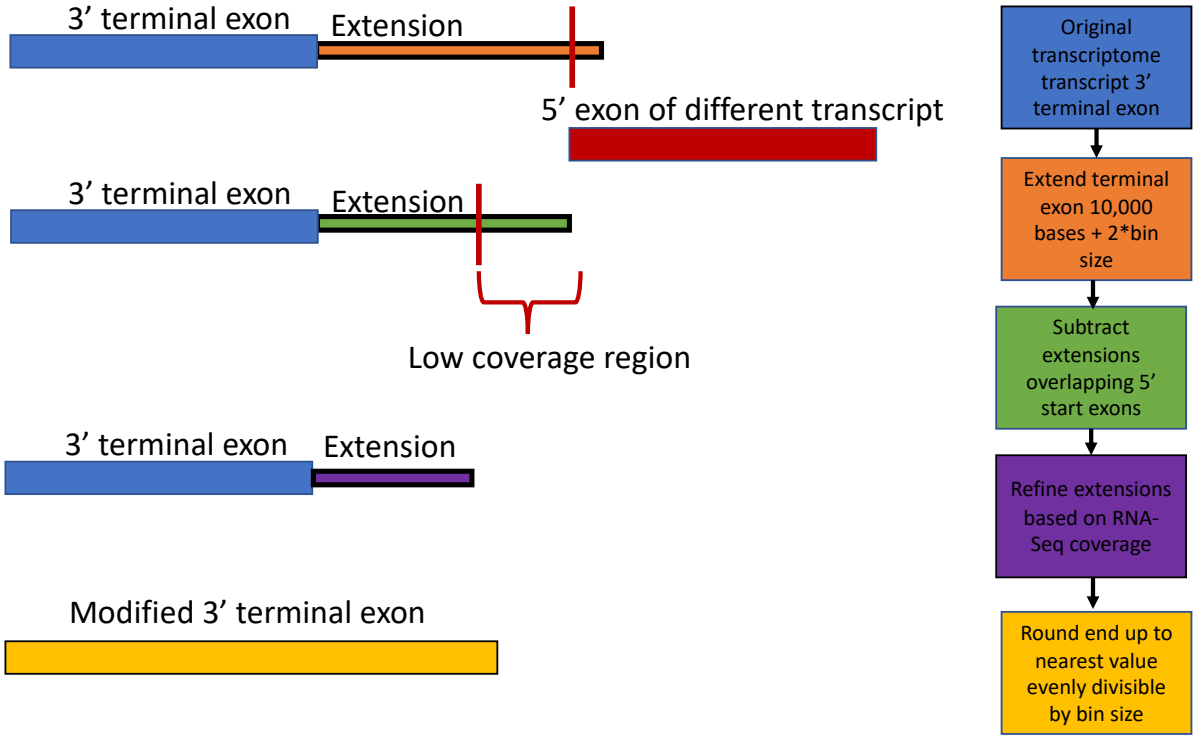

**Supplementary Fig. 4: Simple depiction of the a, intra- and b, inter-bin comparisons used to engineer RNA sequencing features.** For the **a,** intra-bin comparison, the bin of interest (default 100 bases) was divided into three roughly equally sized regions – R1, R2, and R3 – representing the beginning, middle, and end region of the bin, respectively. For the **b,** inter-bin comparisons, the bin of interest was considered R2, and the 100 bases upstream and downstream the bin were considered R1 and R3, respectively.



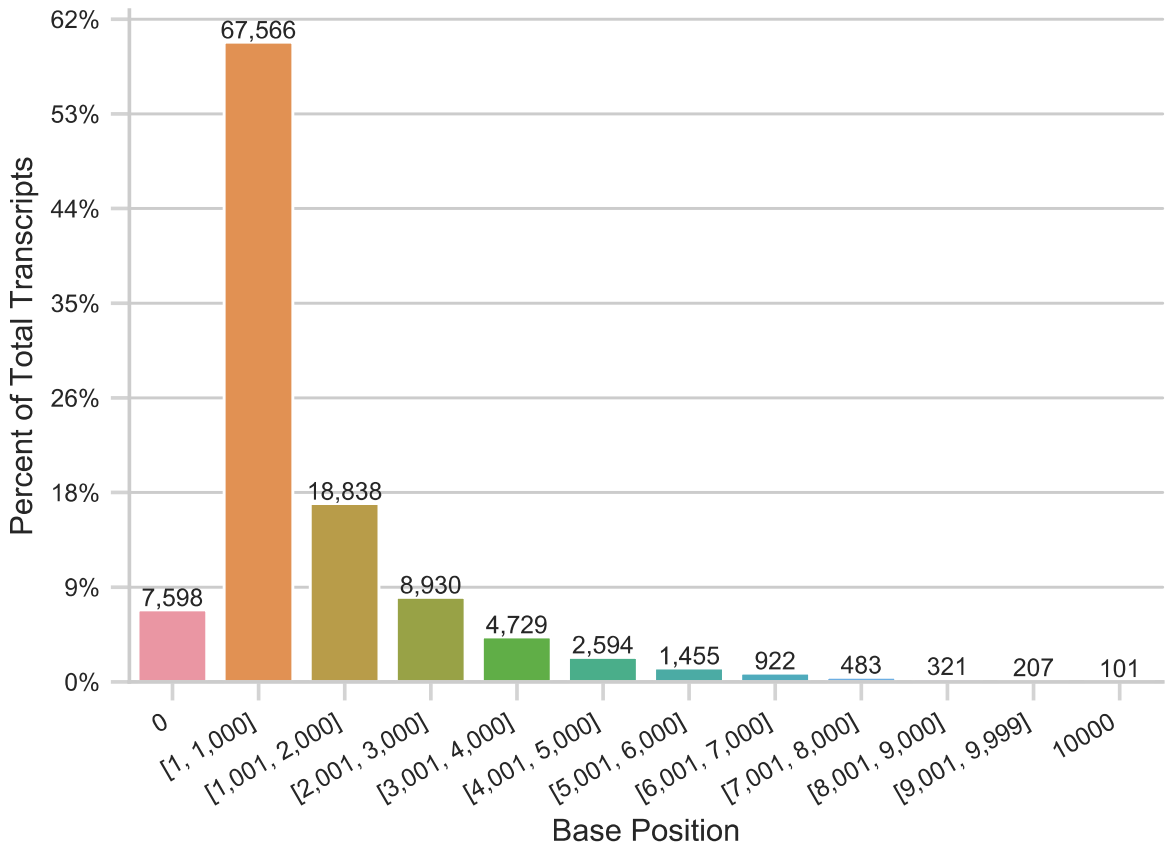

**Supplementary Fig. 5: The data processing pipeline used by aptardi prior to machine learning.** The 3' terminal exons of input transcripts are processed by aptardi (yellow) followed by feature extraction (blue).




**Supplementary Fig. 6: Flowchart depicting the differential expression analysis between the two inbred rat strains, BNLx and SHR.** Yellow boxes denote raw data, green boxes denote data that we generated, and blue boxes denote the two transcriptomes separately subjected to RSEM (RNA-Seq by Expectation Maximization) for comparison.




**Supplementary Fig. 7: Graphical depiction of the transcript processing steps.** The 3’ terminal exon of a transcript derived from the original transcriptome (blue) is first extended 10,000 bases plus two times the bin size (orange). If the extension overlapped the 5’ exon of a neighboring transcript on the same strand (red), the extension was reduced to remove the overlap (green). Next the RNA-sequencing (RNA-Seq) coverage at single nucleotide resolution was used to shorten the 3’ terminal exon to only include regions with detectable coverage relative to the start of the 3’ terminal exon (purple). Finally, the 3’ terminal exon was rounded up to the nearest value evenly divisible by the bin size for compatibility with machine learning (yellow).




**Supplementary Fig. 8: Results from truncating modified 3' terminal exon extensions based on transcript coverage.** Transcripts were shortened based on coverage as described in Transcript processing section of Methods. The base position relative to the start of the terminal exon is given on the x-axis. Over half the modified 3' terminal exons were shortened to <= 1,000 bases. A base position value of zero indicates the transcript was removed entirely because its modified 3' terminal exon did not meet the minimum coverage requirements. Data shown are from the Human Brain Reference dataset.

**Supplementary Table 1: Datasets used to evaluate aptardi.**

| | RNA | | | | DNA Source | True Polyadenyation Sites Source |
|---|---|---|---|---|---|---|
| Dataset | Source | Read Length | Stranded? | # Reads | | |
| HBR | Human Brain Reference | 100 | Yes | 115,926,448 | hg38/GRCh38 | PolyA-Seq Human Brain Reference Total RNA |
| 2nd HBR | Human Brain Reference | 75 | Yes | 139,851,362 | hg38/GRCh38 | PolyA-Seq Human Brain Reference Total RNA |
| UHR | Universal Human Reference | 75 | Yes | 145,513,666 | hg38/GRCh38 | PolyA-Seq Universal Human Reference Total RNA |
| SHR | SHR Inbred Rat Brain | 100 | No | 111,812,107 | SHR Strain Specific | PolyA-Seq Sprague Dawley Total RNA |
| BNLx | BNLx Inbred Rat Brain | 100 | No | 74,863,513 | BNLx Strain Specific | PolyA-Seq Sprague Dawley Total RNA |

**Supplementary Table 2: Comparison of the positive predictive value (PPV) and number of polyadenylation (polyA) sites annotated between the original transcriptome, aptardi modified transcriptome, TAPAS [207], and APARENT [399] at different base distance cutoffs and utilizing different polyA site annotation databases.**

| | Source of PolyA Sites | Base Distance Cutoff | True Positives | False Positives | PPV | Number of PolyA Sites Annotated |
|---|---|---|---|---|---|---|
| Original Transcriptome | HBR PolyA-Seq | 100 | 39,842 | 74,081 | 0.35 | 23,685 |
| Aptardi Modified Transcriptome | | | 62,688 | 79,088 | 0.44 | 29,327 |
| TAPAS | | | 22,804 | 51,810 | 0.31 | 25,180 |
| APARENT | | | 33,213 | 238,883 | 0.14 | 27,999 |
| Original Transcriptome | | 50 | 35,731 | 78,192 | 0.31 | 19,511 |
| Aptardi Modified Transcriptome | | | 49,025 | 92,751 | 0.35 | 23,192 |
| TAPAS | | | 18,357 | 56,257 | 0.25 | 19,064 |
| APARENT | | | 23,562 | 248,534 | 0.09 | 22,153 |
| Original Transcriptome | | 25 | 30,761 | 78,192 | 0.28 | 16,236 |
| Aptardi Modified Transcriptome | | | 38,044 | 103,732 | 0.27 | 18,226 |
| TAPAS | | | 14,303 | 60,311 | 0.19 | 14,281 |
| APARENT | | | 19,560 | 252,536 | 0.08 | 18,371 |
| Original Transcriptome | PolyASite 2.0 | 100 | 51,191 | 62,712 | 0.45 | 45,562 |
| Aptardi Modified Transcriptome | | | 76,277 | 65,452 | 0.54 | 54,925 |
| TAPAS | | | 33,481 | 41,133 | 0.45 | 51,286 |
| APARENT | | | 73,232 | 198,864 | 0.37 | 80,512 |
| Original Transcriptome | | 50 | 44,249 | 69,654 | 0.39 | 31,861 |
| Aptardi Modified Transcriptome | | | 60,900 | 80,829 | 0.43 | 37,842 |
| TAPAS | | | 26,418 | 48,196 | 0.35 | 33,567 |
| APARENT | | | 54,665 | 217,431 | 0.25 | 59,115 |
| Original Transcriptome | | 25 | 36,996 | 76,907 | 0.32 | 21,969 |
| Aptardi Modified Transcriptome | | | 46,973 | 94,756 | 0.33 | 25,218 |
| TAPAS | | | 20,063 | 54,551 | 0.27 | 20,743 |
| APARENT | | | 43,722 | 228,374 | 0.19 | 42,425 |
| Original Transcriptome | PolyA_DB | 100 | 49,648 | 64,255 | 0.44 | 41,893 |
| Aptardi Modified Transcriptome | | | 75,531 | 66,198 | 0.53 | 51,381 |
| TAPAS | | | 31,379 | 43,235 | 0.42 | 46,125 |
| APARENT | | | 63,249 | 208,847 | 0.30 | 66,770 |
| Original Transcriptome | | 50 | 43,960 | 69,943 | 0.39 | 30,717 |
| Aptardi Modified Transcriptome | | | 61,348 | 80,381 | 0.43 | 36,974 |
| TAPAS | | | 25,319 | 49,295 | 0.34 | 31,369 |
| APARENT | | | 47,852 | 224,244 | 0.21 | 50,794 |
| Original Transcriptome | | 25 | 37,670 | 76,233 | 0.33 | 22,340 |
| Aptardi Modified Transcriptome | | | 48,112 | 93,617 | 0.34 | 25,833 |
| TAPAS | | | 19,482 | 55,132 | 0.26 | 20,307 |
| APARENT | | | 40,149 | 231,947 | 0.17 | 38,965 |

A prediction was considered a true positive if it was within the given base distance cutoff of an annotated polyA site. Annotated polyA sites were taken from the human brain reference (HBR) PolyA-Seq data, PolyASite 2.0[404], and PolyA_DB[405]. The original transcriptome was generated from the HBR dataset, and predictions by aptardi, TAPAS, and APARENT were made using these transcript structures. Namely, TAPAS used the HBR RNA-Seq data, APARENT used the hg38/GRCh38 reference human genome, and aptardi used both.

**Supplementary Table 3: RNA sequencing alignment results for mouse tissue analysis.**

| Dataset | Brain | Liver |
|---|---|---|
| # Reads | 15,239,319 | 15,991,252 |
| Overall Genome Alignment Rate | 98.70% | 97.46% |

Reads were aligned to the mm10/GRCm38 mouse reference genome with HISAT2 (v.2.1.0).

**Supplementary Table 4: Few polyadenylation (polyA) sites share a 100 base region with another polyA site**.

| Total # PolyA Sites Captured | # PolyA Sites Sharing 100 Base Bin | # Multi PolyA 100 Base Bins |
|---|---|---|
| 42,977 | 3,625 | 1,807 |

Since aptardi makes predictions in 100 base increments, sites within 100 bases of one another cannot be distinguished. Data shown are from the Human Brain Reference dataset.

**Supplementary Table 5: RNA sequencing alignment results for each sample.**

| Dataset | HBR | 2nd HBR | UHR | BNLx | SHR |
|---|---|---|---|---|---|
| # Reads | 115,926,448 | 139,851,362 | 145,513,666 | 74,863,513 | 111,812,107 |
| Overall Genome Alignment Rate | 96.58% | 95.39% | 95.38% | 96.41% | 96.76% |

Reads were aligned to each sample's respective genome with HISAT2 (v. 2.1.0).

**Supplementary Table 6: RNA sequencing alignment results for the CFIm25 knockdown analysis.**

| Dataset | Control | CFIm25 Knockdown |
|---|---|---|
| # Reads | 164,774,179 | 160,083,915 |
| Overall Genome Alignment Rate | 94.94% | 95.91% |

Reads were aligned to the hg38/GRCh38 human reference genome with HISAT2 (v. 2.1.0).

**Supplementary Table 7: The transcript processing steps increase the number of polyadenylation sites included in aptardi analysis.**

| Transcript Processing Step | # Polyadenylation Sites from Source | | |
|---|---|---|---|
| [redacted] | Transcript Terminal Exon | Transcript Extension | Both a Transcript's Terminal Exon and a Separate Transcript's Extension |
| Terminal Exon + Extension (Original) | 24,640 | 24,107 | 20,707 |
| Subtract Overlapping Starts | 24,635 | 22,356 | 10,336 |
| Truncate Based on Coverage | 20,072 | 6,189 | 8,097 |
| Window (Final) | 20,541 | 7,437 | 8,390 |

The number of unique polyadenylation sites captured at each step is shown, along with the category from which the site was derived.

**Supplementary Table 8: Summary of engineered DNA sequence features.**

| DNA Sequence Element | Nucleotide String(s) | Window Size | Region Probed, if PAS Present (Relative to PAS) | Region Probed, if PAS not Present (Relative to Bin Start (for Start), Bin End (for End)) | Frequency of String Required for Enrichment (>=) |
|---|---|---|---|---|---|
| Distal downstream G-rich region | >=5 G's | 6 | +43 to +143 (or end*) | +30 to end* | 0.0585 |
| Proximal downstream T-rich region | TTT | 3 | +13 to +76 | +10 to +40 | 0.125 |
| Proximal downstream GT/TG-rich region | GT & TG | 2 | | | 0.25 |
| Proximal downstream GTGT/TGTG-rich region | GTGT & TGTG | 4 | | | 0.0469 |
| Intermediate T-rich region | T | 1 | +6 to +36 | -36 to 0 | 0.375 |
| Upstream T-rich region | T | 1 | -50 to 0 | -86 to -7 | 0.375 |
| Upstream TGTA/TATA-rich region | TGTA & TATA | 4 | -40 to 0 | -76 to -7 | 0.0469 |
| AT-rich region | AT | 2 | -93 (or start*) to +142 (or end*) | Start* to end* | 0.125 |

*Start = 100 bases upstream bin start, end = 100 bases downstream bin end

**Supplementary Information**

**Transcript processing**

Modified 3' terminal exons were refined using an approach similar to that described by Ye et al. [206] and Miura et al. [384] as follows. If the average coverage of the first X bases (X = bin size) of the modified 3' terminal exon was less than 10% of the entire transcript's average coverage and/or the modified 3' terminal exon was not at least three times the bin size (default 100 bases), the transcript was removed. Otherwise the transcript's modified 3' terminal exon was scanned 5' to 3' using a sliding window equal to the bin size until the following metrics were less than 5% of the average coverage of the first bases equal to the bin size of the modified 3' terminal exon: 1) 80% of the bases in the current bin, 2) the average coverage of the previous bin, 3) the average coverage of the subsequent bin, and 4) the coverage of the current base (i.e. first base in the current bin). This strategy is robust to poor local coverage that can occur in RNA-Seq data (e.g. GC bias). The base that meets these criteria defines the end of the modified 3' terminal exon for the transcript, i.e. this base is not considered a transcript stop site but rather defines the 3' end of the region that will be explored by aptardi. For compatibility with machine learning, where predictions are made on a set bin size (i.e. 100 base bins as the default), each modified 3' terminal exon was rounded up to the nearest value evenly divisible by the bin size at the 3' end. Supplementary Fig. 7 graphically depicts these transcript processing steps. Note that since the coverage of the current and subsequent bins are used when refining modified 3' terminal exons, the longest possible 3' modified terminal exon is two times the bin size less than its total length.

To evaluate the impact of transcript processing on the original transcriptome fed to aptardi, we first ascertained the number of unique polyadenylation (polyA) sites captured at each

step and further determined from which of the following three categories each was derived: 1) the original reconstruction terminal exon, 2) the extension step, or 3) both (1) and (2) as a result of overlaps (Supplementary Table 8). The extension step doubled the number of polyA sites captured. After subtracting extensions overlapping a neighboring transcript's start, the number of polyA sites in (3) was halved. This suggests the extension step resulted in extensions long enough to encompass entire neighboring transcripts, supporting the need to subtract overlap. Shrinking extension length once again based on transcript coverage (see Transcript processing section in Methods) reduced the number of polyA sites captured in (2) by more than a third and removed 7,598 transcripts from analysis (Supplementary Fig. 8). This decrease is large but likely necessary to ensure polyA sites captured by a given transcript plus extension confidently belong to that extension and is being expressed. Overall, more than 7,000 novel transcript stop sites were included in aptardi analysis though transcript processing.

**DNA sequence features**

All DNA sequence features were encoded as binary indicators to indicate presence (1) or absence (-1) in each bin (default 100 bases).

For each of the four polyadenylation signals (PAS's) – 1) AATAAA, 2) ATTAAA, 3) AGTAAA and any of 4) AAGAAA, AAAAAG, AATACA, TATAAA, GATAAA, AATATA, CATAAA, AATAGA – a sliding six base window was scanned from -35 bases upsteam the bin start to -7 bases upstream the base end in single nucleotide increments. If any single hexamer matched the given PAS, it was encoded 1, otherwise -1.

In general, the 100 bases upstream and downstream the bin, as well as the bin itself (300 bases total for the default 100 base bin size) were used for the DNA sequence elements features; however, the specific region examined for each DNA sequence element varied by the given

feature and whether a PAS was present. If more than one PAS was present, the PAS that dictated the region probed was first by priority in the order listed above, i.e. if AATAAA and ATTAAA were present, the location of AATAAA was used, and next by the first occurrence of the location, i.e. if AATAAA was present multiple times, the location of the 5' most signal was used.

The following DNA sequence elements were evaluated: 5) a distal downstream G-rich region, a proximal downstream region enriched in 6) T, 7) GT/TG, and 8) GTGT/TGTG, an intermediate 9) T-rich region, an upstream region enriched in 10) T and 11) TGTA/TATA, and a surrounding 12) AT-rich region. A similar sliding window strategy was utilized, but here the number of windows matching the element to the number of windows not matching the element, i.e., its frequency, was compared to an enrichment threshold value to determine if the given element was considered enriched, encoded 1, or not, encoded (-1). Enrichment thresholds varied across elements. Supplementary Table 9 summarizes the DNA sequence features.

**RNA sequencing features**

RNA-Seq features were engineered by defining an upstream region (R1), middle region (R2), and downstream region (R3) for each of the following: 1) intra- and 2) inter-bin. For intra-bin, the 100 base bin was divided into 34, 33, and 33 bases 5' to 3'. For inter-bin, the 100 bases 5' the 100 base bin, the bin itself, and the 100 bases 3' the bin served as R1, R2, and R3, respectively. The median coverage values of the regions were combined in seven ways for each the intra- and inter-bin to give 14 features:

1) R1-R2
2) R2-R3
3) R1/(R1+R2+R3)

4) R2/(R1+R2+R3)

5) R3/(R1+R2+R3)

6) R2/(R1+R3)

7) R3/(R1+R3)

Note that if the denominator equaled zero, the feature was given a zero.

Supplementary Data 1.

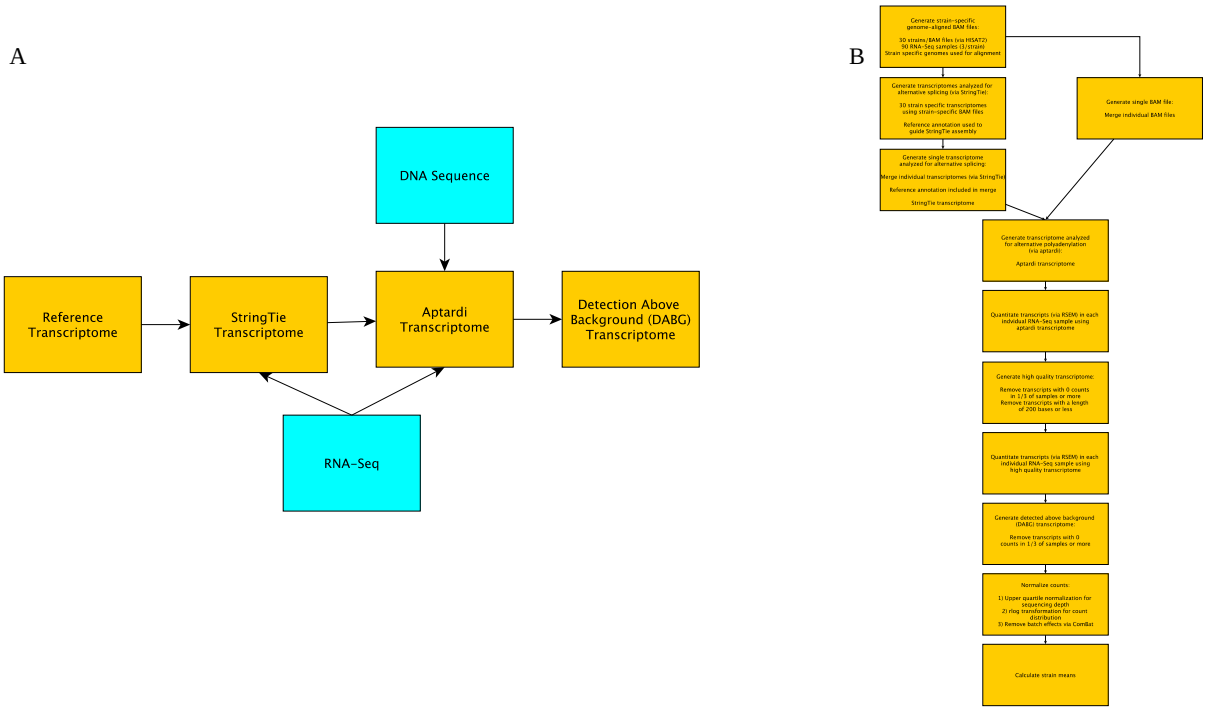
| Reference Ensembl Gene Name | Reference Ensembl Gene | Reference Ensembl Transcript | aptamB Transcript | StringTie Transcript | aptamB Base Mean | StringTie Base Mean | aptamB log2-Fold Change | StringTie log2-Fold Change | aptamB p-value | StringTie p-value | aptamB Adjusted p-value | StringTie Adjusted p-value | aptamB Standard Error of the log2-Fold Change estimate | StringTie Standard Error of the log2-Fold Change estimate | aptamB Wald statistic | StringTie Wald statistic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gng12 | ENSRNOG00000005031 | ENSRNOT00000075400 | STRG.17426.2.2 | STRG.17426.2 | 526.41 | 73.9 | -1.62 | -0.03 | 6.44E-06 | 9.53E-01 | 1.09E-02 | 9.95E-01 | 0.13 | 0.45 | -12.51 | -0.06 |
| RGD1559441 | ENSRNOG00000024953 | ENSRNOT00000042810 | STRG.19724.5.1 | STRG.19724.5 | 680.37 | 97.14 | -1.48 | -0.52 | 8.75E-20 | 2.45E-01 | 4.61E-17 | 8.25E-01 | 0.16 | 0.45 | -9.1 | -1.16 |
| Ppm3 | ENSRNOG00000011425 | ENSRNOT00000076255 | STRG.20656.3.1 | STRG.20656.3 | 915.81 | 49.05 | -12.11 | -0.21 | 1.51E-19 | 9.43E-01 | 7.48E-17 | 9.94E-01 | 1.34 | 2.89 | -9.04 | -0.07 |
| Grb | ENSRNOG00000010199 | ENSRNOT00000014170 | STRG.9887.1 | STRG.9887.1 | 808.3 | 4122.37 | -1.36 | -0.45 | 5.05E-19 | 4.81E-03 | 2.29E-16 | 1.05E-01 | 0.15 | 0.16 | -8.91 | -2.82 |
| Khsrp | ENSRNOG00000047628 | ENSRNOT00000077653 | STRG.33444.2 | STRG.33444.2 | 159.99 | 840.4 | 10.72 | 0.51 | 1.60E-16 | 1.87E-03 | 5.20E-14 | 5.12E-02 | 1.3 | 0.16 | 8.25 | 3.11 |
| Hnrnpa3 | ENSRNOG00000010474 | ENSRNOT00000016137 | STRG.17838.1.1 | STRG.17838.1 | 233.84 | 3757.46 | -11.68 | 0.1 | 2.44E-16 | 5.09E-01 | 7.70E-14 | 9.41E-01 | 1.42 | 0.16 | -8.2 | 0.66 |
| Trip12 | ENSRNOG00000010963 | ENSRNOT00000093245 | STRG.35041.2.1 | STRG.35041.2 | 171.68 | 1490.42 | -9.46 | 0.43 | 1.32E-15 | 2.76E-01 | 3.76E-13 | 8.48E-01 | 1.18 | 0.39 | -7.99 | 1.09 |
| Ikzf3 | ENSRNOG00000044751 | ENSRNOT00000068321 | STRG.24972.1.2 | STRG.24972.1 | 127.12 | 758.24 | 10.11 | -0.11 | 3.46E-15 | 4.95E-01 | 8.99E-13 | 9.39E-01 | 1.28 | 0.16 | 7.87 | -0.68 |
| Ppm1k | ENSRNOG00000006893 | ENSRNOT00000008785 | STRG.17338.1.1 | STRG.17338.3 | 164.24 | 73.3 | -11.17 | 1.57 | 8.56E-15 | 4.42E-02 | 2.11E-12 | 4.28E-01 | 1.44 | 0.78 | -7.76 | 2.01 |
| Tpi1lbp1 | ENSRNOG00000013837 | ENSRNOT00000013023 | STRG.14036.1 | STRG.14036.1 | 145.98 | 1154.4 | 10.31 | 0.27 | 1.29E-14 | 4.77E-01 | 2.98E-12 | 9.35E-01 | 1.34 | 0.37 | 7.71 | 0.71 |
| Fam3 | ENSRNOG00000015258 | ENSRNOT00000083033 | STRG.29714.1.4 | STRG.29714.1 | 3289.22 | 531.05 | -9.66 | -0.43 | 2.68E-14 | 1.67E-01 | 5.78E-12 | 7.49E-01 | 0.09 | 0.33 | -7.61 | -1.38 |
| Ube4a | ENSRNOG00000026833 | ENSRNOT00000021321 | STRG.30825.4 | STRG.30825.4 | 136.21 | 1627.03 | 10.21 | 0.02 | 1.03E-13 | 9.10E-01 | 2.05E-11 | 9.94E-01 | 1.37 | 0.17 | 7.44 | 0.11 |
| Pdkb | ENSRNOG00000021024 | ENSRNOT00000040501 | STRG.10341.3.1 | STRG.10341.3 | 71.97 | 875.24 | -9.98 | -0.09 | 9.40E-13 | 6.75E-01 | 1.60E-10 | 9.70E-01 | 1.4 | 0.2 | -7.14 | -0.42 |
| Mmadhc | ENSRNOG00000004740 | ENSRNOT00000066058 | STRG.12652.1.1 | STRG.12652.1 | 59.08 | 249.22 | -7.06 | -0.54 | 2.81E-12 | 1.60E-01 | 4.56E-10 | 7.41E-01 | 1.01 | 0.39 | -6.99 | -1.4 |
| Dut | ENSRNOG00000007221 | ENSRNOT00000009549 | STRG.14123.1 | STRG.14123.1 | 63.52 | 237.14 | 9.11 | -0.87 | 1.58E-11 | 5.02E-03 | 2.36E-09 | 1.08E-01 | 1.35 | 0.31 | 6.74 | -2.81 |
| Tmem88b | ENSRNOG00000036869 | ENSRNOT00000055343 | STRG.23109.1.2 | STRG.23109.1 | 575.47 | 460.31 | -2.46 | -0.79 | 1.78E-11 | 6.30E-02 | 2.64E-09 | 5.14E-01 | 0.37 | 0.43 | -6.72 | -1.86 |
| Zfp915a | ENSRNOG00000048577 | ENSRNOT00000082664 | STRG.26763.6 | STRG.26763.6 | 54.14 | 248.67 | 8.88 | 0.49 | 3.42E-11 | 2.18E-02 | 4.91E-09 | 2.88E-01 | 1.34 | 0.21 | 6.63 | 2.29 |
| Lrrdd1 | ENSRNOG00000040684 | ENSRNOT00000068687 | STRG.24058.2.2 | STRG.24058.2 | 928.19 | 1552.4 | -2.33 | -0.03 | 4.71E-11 | 8.85E-01 | 6.62E-09 | 9.92E-01 | 0.35 | 0.18 | -6.58 | -0.14 |
| Ndufa4 | ENSRNOG00000005512 | ENSRNOT00000007567 | STRG.16399.3.2 | STRG.16399.3 | 161.35 | 3869.52 | -8.09 | -0.24 | 8.76E-11 | 3.69E-01 | 1.17E-08 | 8.98E-01 | 1.25 | 0.27 | -6.49 | -0.9 |
| Zfp317 | ENSRNOG00000036684 | ENSRNOT00000078593 | STRG.30210.8.1 | STRG.30210.8 | 48 | 88 | -9.39 | 0.53 | 2.25E-10 | 2.13E-01 | 2.80E-08 | 7.99E-01 | 1.48 | 0.43 | -6.34 | 1.25 |
| Tmem151b | ENSRNOG00000013958 | ENSRNOT00000013706 | STRG.33715.1 | STRG.33715.1 | 97.4 | 730.01 | 9.73 | 0.44 | 3.76E-10 | 1.90E-01 | 4.44E-08 | 7.80E-01 | 1.55 | 0.33 | 6.26 | 1.31 |
| Actr10 | ENSRNOG00000007504 | ENSRNOT00000010200 | STRG.24931.1 | STRG.24931.1 | 70.43 | 1300.9 | -9.95 | -0.22 | 4.28E-10 | 2.14E-01 | 5.00E-08 | 8.01E-01 | 1.59 | 0.17 | -6.24 | -1.24 |
| Rnf13 | ENSRNOG00000029209 | ENSRNOT00000068606 | STRG.9331.2.1 | STRG.9331.2 | 121.98 | 1372.6 | 8.12 | -0.13 | 1.02E-09 | 6.73E-01 | 1.15E-07 | 9.70E-01 | 1.33 | 0.3 | 6.11 | -0.42 |
| Fam3 | ENSRNOG00000015258 | ENSRNOT00000083033 | STRG.29714.1.3 | STRG.29714.1 | 68.56 | 531.05 | -9.91 | -0.43 | 1.56E-09 | 1.67E-01 | 1.70E-07 | 7.49E-01 | 1.64 | 0.31 | -6.04 | -1.38 |
| Ppm3 | ENSRNOG00000011425 | ENSRNOT00000076705 | STRG.20656.3.2 | STRG.20656.3 | 156.86 | 49.05 | -11.1 | -0.21 | 1.98E-09 | 9.43E-01 | 2.09E-07 | 9.94E-01 | 1.85 | 2.89 | -6 | -0.07 |
| Epb41l4b | ENSRNOG00000056550 | ENSRNOT00000090366 | STRG.20652.1.1 | STRG.20652.1 | 37.56 | 70.7 | 8.36 | -3.23 | 2.24E-09 | 2.23E-01 | 2.32E-07 | 8.08E-01 | 1.4 | 2.65 | 5.98 | -1.22 |
| Copg2 | ENSRNOG00000031014 | ENSRNOT00000033706 | STRG.16731.1 | STRG.16731.1 | 70.77 | 1497.94 | 9.27 | -0.03 | 4.60E-09 | 8.87E-01 | 4.63E-07 | 9.92E-01 | 1.58 | 0.2 | 5.86 | -0.14 |
| Topors | ENSRNOG00000006485 | ENSRNOT00000006081 | STRG.20183.3.1 | STRG.20183.3 | 55.86 | 711.48 | 8.92 | -0.04 | 5.42E-09 | 8.12E-01 | 5.39E-07 | 9.86E-01 | 1.53 | 0.15 | 5.83 | -0.24 |
| Copg2 | ENSRNOG00000031014 | ENSRNOT00000033706 | STRG.16731.1.2 | STRG.16731.1 | 49.97 | 1497.94 | -9.45 | -0.03 | 1.21E-08 | 8.87E-01 | 1.13E-06 | 9.92E-01 | 1.66 | 0.2 | -5.7 | -0.14 |
| Jkamp | ENSRNOG00000044751 | ENSRNOT00000068321 | STRG.24972.1 | STRG.24972.1 | 32.17 | 758.24 | 8.13 | -0.11 | 2.40E-08 | 4.95E-01 | 2.14E-06 | 9.39E-01 | 1.46 | 0.16 | 5.58 | -0.68 |
| Fgf11op2 | ENSRNOG00000001811 | ENSRNOT00000042270 | STRG.19159.1.1 | STRG.19159.1 | 44.27 | 351.79 | -9.28 | 0.02 | 2.74E-08 | 9.46E-01 | 2.40E-06 | 9.94E-01 | 1.67 | 0.28 | -5.56 | 0.07 |
| Aplar | ENSRNOG00000024492 | ENSRNOT00000055607 | STRG.11251.1.1 | STRG.11251.1 | 65.25 | 141.4 | 9.15 | 0.19 | 2.88E-08 | 5.89E-01 | 2.51E-06 | 9.80E-01 | 1.65 | 0.35 | 5.55 | 0.54 |
| Fjx1 | ENSRNOG00000015528 | ENSRNOT00000021258 | STRG.35480.4.3 | STRG.35480.4 | 657.97 | 817.42 | -5.24 | 1.02 | 1.58E-07 | 7.83E-01 | 1.20E-05 | 9.81E-01 | 1 | 3.73 | -5.24 | 0.28 |
| Dpf3b | ENSRNOG00000030771 | ENSRNOT00000063874 | STRG.24384.1.1 | STRG.24384.1 | 566.6 | 1939.72 | 1.55 | 0.42 | 1.98E-07 | 4.16E-01 | 1.48E-05 | 9.17E-01 | 0.3 | 0.52 | 5.2 | 0.81 |
| Jkamp | ENSRNOG00000044751 | ENSRNOT00000068321 | STRG.24972.1.1 | STRG.24972.1 | 858.16 | 758.24 | -0.74 | -0.11 | 2.07E-07 | 4.95E-01 | 1.54E-05 | 9.39E-01 | 0.14 | 0.16 | -5.19 | -0.68 |
| Rab11a | ENSRNOG00000011302 | ENSRNOT00000015598 | STRG.31436.2.3 | STRG.31436.2 | 66.71 | 1838.84 | 9.18 | -0.1 | 3.03E-07 | 2.69E-01 | 2.16E-05 | 8.44E-01 | 1.79 | 0.09 | 5.12 | -1.11 |
| Satb1 | ENSRNOG00000012942 | ENSRNOT00000017556 | STRG.33300.4.4 | STRG.33300.4 | 150.5 | 934.64 | -8.08 | 0.45 | 7.10E-07 | 4.85E-01 | 4.65E-05 | 9.36E-01 | 1.63 | 0.64 | -4.96 | 0.7 |
| Pcsk2 | ENSRNOG00000005438 | ENSRNOT00000007249 | STRG.14596.1 | STRG.14596.1 | 3388.64 | 6811.69 | -0.84 | -0.42 | 8.33E-07 | 9.68E-02 | 5.37E-05 | 6.23E-01 | 0.17 | 0.25 | -4.93 | -1.66 |
| Thy1 | ENSRNOG00000006604 | ENSRNOT00000008685 | STRG.30776.1.1 | STRG.30776.1 | 17688.34 | 7636.16 | -0.49 | -0.32 | 1.29E-06 | 5.23E-02 | 8.04E-05 | 4.67E-01 | 0.1 | 0.16 | -4.84 | -1.94 |
| Fam3 | ENSRNOG00000015258 | ENSRNOT00000083033 | STRG.29714.1.2 | STRG.29714.1 | 120.08 | 531.05 | -10.72 | -0.43 | 1.74E-06 | 1.67E-01 | 0.000105485 | 7.49E-01 | 2.24 | 0.31 | -4.78 | -1.38 |
| Lrrdd1 | ENSRNOG00000040684 | ENSRNOT00000068687 | STRG.24058.2 | STRG.24058.2 | 460.68 | 1552.4 | 5.24 | -0.03 | 4.50E-06 | 8.85E-01 | 0.000249819 | 9.92E-01 | 1.14 | 0.18 | 4.59 | -0.14 |
| Sts12 | ENSRNOG00000011804 | ENSRNOT00000017227 | STRG.22270.5.3 | STRG.22270.5 | 471.12 | 2732.59 | -1.23 | -0.01 | 1.16E-05 | 9.43E-01 | 0.000599946 | 9.94E-01 | 0.28 | 0.11 | -4.39 | -0.07 |
| Ufm1 | ENSRNOG00000038176 | ENSRNOT00000043116 | STRG.9247.1 | STRG.9247.1 | 77.55 | 500.41 | -4.69 | -0.26 | 1.26E-05 | 2.59E-01 | 0.000640838 | 8.38E-01 | 1.07 | 0.23 | -4.37 | -1.13 |
| Ireig1 | ENSRNOG00000006859 | ENSRNOT00000009233 | STRG.15800.1.1 | STRG.15800.1 | 1514.69 | 1686.01 | -0.63 | -0.21 | 2.70E-05 | 2.88E-01 | 0.001301432 | 8.58E-01 | 0.15 | 0.19 | -4.2 | -1.06 |
| Sli | ENSRNOG00000053340 | ENSRNOT00000080591 | STRG.23071.1 | STRG.23071.1 | 222.58 | 692.23 | 1.72 | 0.87 | 3.20E-05 | 1.08E-03 | 0.001511097 | 3.27E-02 | 0.41 | 0.26 | 4.16 | 3.27 |
| Arhgap5 | ENSRNOG00000044696 | ENSRNOT00000068378 | STRG.24617.1.1 | STRG.24617.1 | 166.04 | 1195.26 | -11.19 | 0.52 | 3.49E-05 | 1.80E-01 | 0.001640115 | 7.67E-01 | 2.7 | 0.39 | -4.14 | 1.34 |
| Abcg2 | ENSRNOG00000007041 | ENSRNOT00000009546 | STRG.17339.1 | STRG.17339.1 | 43.4 | 65.92 | 2.3 | 1.83 | 3.53E-05 | 3.82E-03 | 0.001655341 | 8.93E-02 | 0.56 | 0.63 | 4.14 | 2.89 |
| Fgf11op2 | ENSRNOG00000001811 | ENSRNOT00000002476 | STRG.19159.2 | STRG.19159.2 | 611.02 | 937.92 | -0.88 | -0.2 | 4.60E-05 | 5.67E-01 | 0.002080742 | 9.55E-01 | 0.22 | 0.35 | -4.08 | -0.57 |
| Rbsn | ENSRNOG00000011391 | ENSRNOT00000015602 | STRG.17975.1.1 | STRG.17975.1 | 730.44 | 1117.37 | -0.65 | 0.18 | 4.69E-05 | 3.49E-01 | 0.002115485 | 8.91E-01 | 0.16 | 0.2 | -4.07 | 0.94 |
| Pdk1 | ENSRNOG00000001517 | ENSRNOT00000002073 | STRG.13027.1 | STRG.13027.1 | 178.43 | 236.04 | 0.94 | 0.67 | 7.01E-05 | 2.44E-02 | 0.003031391 | 3.09E-01 | 0.24 | 0.3 | 3.98 | 2.25 |
| Dtd | ENSRNOG00000006364 | ENSRNOT00000008980 | STRG.24274.1.3 | STRG.24274.1 | 526.47 | 2573.81 | -0.92 | -0.2 | 7.07E-05 | 7.99E-02 | 0.003051628 | 5.76E-01 | 0.23 | 0.12 | -3.97 | -1.75 |
| Opcml | ENSRNOG00000023809 | ENSRNOT00000042890 | STRG.30471.3.4 | STRG.30471.3 | 102.77 | 244.21 | -4.02 | 2.74 | 7.69E-05 | 4.62E-01 | 0.003273153 | 9.31E-01 | 1.52 | 3.73 | -3.95 | 0.74 |
| Nudcd1 | ENSRNOG00000024929 | ENSRNOT00000006685 | STRG.28266.1 | STRG.28266.1 | 31.82 | 177.21 | 3.09 | 0.67 | 9.02E-05 | 2.95E-02 | 0.003751823 | 3.45E-01 | 0.79 | 0.31 | 3.92 | 2.18 |
| P2ry12 | ENSRNOG00000013902 | ENSRNOT00000018600 | STRG.9388.2 | STRG.9388.2 | 176.27 | 339.19 | 0.85 | 0.16 | 9.40E-05 | 4.27E-01 | 0.003886089 | 9.22E-01 | 0.22 | 0.2 | 3.91 | 0.8 |
| Gpatch11 | ENSRNOG00000025317 | ENSRNOT00000034919 | STRG.23160.1 | STRG.23160.1 | 97.02 | 498.58 | -2.22 | 0.3 | 9.85E-05 | 8.63E-02 | 0.00404911 | 5.92E-01 | 0.57 | 0.17 | -3.89 | 1.72 |
| Ptn | ENSRNOG00000011946 | ENSRNOT00000016088 | STRG.16835.2.1 | STRG.16835.2 | 961.97 | 1635.03 | 0.52 | -0.58 | 0.000127383 | 2.27E-01 | 0.005019146 | 8.10E-01 | 0.14 | 0.48 | 3.83 | -1.21 |
| Dnm1l | ENSRNOG00000039859 | ENSRNOT00000054932 | STRG.30262.1.1 | STRG.30262.1 | 684.56 | 487.92 | 1.08 | 0.23 | 0.000134966 | 2.78E-01 | 0.005272685 | 8.49E-01 | 0.28 | 0.21 | 3.82 | 1.08 |
| Khdc2 | ENSRNOG00000040474 | ENSRNOT00000046223 | STRG.24870.2.1 | STRG.24870.2 | 1156.1 | 1452.15 | -0.59 | -0.04 | 0.000154861 | 6.98E-01 | 0.005968503 | 9.73E-01 | 0.16 | 0.11 | -3.78 | -0.39 |
| Mdh1 | ENSRNOG00000018895 | ENSRNOT00000021564 | STRG.29151.1 | STRG.29151.1 | 217.13 | 368.55 | 1.03 | 0.49 | 0.000167789 | 1.28E-01 | 0.006324457 | 6.88E-01 | 0.27 | 0.32 | 3.76 | 1.52 |
| Sf3b1 | ENSRNOG00000013516 | ENSRNOT00000019126 | STRG.34429.2 | STRG.34429.2 | 280.17 | 3114.6 | 1.32 | 0.23 | 0.000185044 | 6.93E-02 | 0.00683933 | 5.37E-01 | 0.35 | 0.12 | 3.74 | 1.82 |
| Srsf6 | ENSRNOG00000006380 | ENSRNOT00000008427 | STRG.15204.1.1 | STRG.15204.1 | 582.29 | 1009.27 | 0.74 | 0.32 | 0.00019023 | 1.90E-01 | 0.006942423 | 7.80E-01 | 0.2 | 0.25 | 3.73 | 1.31 |
| Zdhhc22 | ENSRNOG00000011285 | ENSRNOT00000015046 | STRG.25421.1.1 | STRG.25421.1 | 822.49 | 216.19 | -1.6 | 0.68 | 0.000221573 | 4.92E-01 | 0.007972792 | 9.39E-01 | 0.43 | 0.7 | -3.69 | 0.69 |
| Rpk1 | ENSRNOG00000008051 | ENSRNOT00000008829 | STRG.25626.1.1 | STRG.25626.1 | 82.4 | 239.64 | -10.18 | 0.01 | 0.000267494 | 9.67E-01 | 0.00935851 | 9.97E-01 | 2.79 | 0.31 | -3.64 | 0.04 |
| Fbln1 | ENSRNOG00000018903 | ENSRNOT00000073057 | STRG.7508.2.1 | STRG.7508.2 | 79.58 | 139.73 | -10.13 | 0.52 | 0.000309594 | 4.90E-01 | 0.010657475 | 9.38E-01 | 2.81 | 0.75 | -3.61 | 0.69 |
| Ndufa10l1 | ENSRNOG00000016470 | ENSRNOT00000021089 | STRG.35233.1 | STRG.35233.1 | 947.32 | 2547.87 | -0.65 | -0.43 | 0.000321498 | 1.26E-03 | 0.010961617 | 3.73E-02 | 0.18 | 0.13 | -3.6 | -3.22 |
| Card1 | ENSRNOG00000007834 | ENSRNOT00000010720 | STRG.27721.2.1 | STRG.27721.2 | 299.71 | 2593.58 | -2.16 | 0.13 | 0.000324251 | 5.81E-01 | 0.011043779 | 9.58E-01 | 0.6 | 0.23 | -3.6 | 0.55 |
| Cacna2d1 | ENSRNOG00000039531 | ENSRNOT00000043914 | STRG.16084.1 | STRG.16084.1 | 142.7 | 142.27 | 10.28 | 10.27 | 0.000350573 | 6.60E-03 | 0.011753525 | 1.59E-01 | 2.88 | 3.91 | 3.57 | 2.63 |
| Unc79 | ENSRNOG00000008728 | ENSRNOT00000011775 | STRG.25642.4 | STRG.25642.4 | 141.12 | 1136.02 | 10.26 | 0.49 | 0.000359659 | 5.10E-03 | 0.011970762 | 1.10E-01 | 2.88 | 0.18 | 3.57 | 2.8 |
| Sfxn5 | ENSRNOG00000037871 | ENSRNOT00000057537 | STRG.17725.1 | STRG.17725.1 | 135.77 | 1249.98 | 10.21 | 0.98 | 0.000390751 | 4.79E-03 | 0.012832413 | 1.05E-01 | 2.88 | 0.35 | 3.55 | 2.82 |
| Pbp2 | ENSRNOG00000010827 | ENSRNOT00000010036 | STRG.11045.1.2 | STRG.11045.1 | 614.76 | 884.16 | -1.37 | 0.24 | 0.000421981 | 6.02E-02 | 0.01369074 | 5.03E-01 | 0.39 | 0.13 | -3.53 | 1.88 |
| Casp8 | ENSRNOG00000012944 | ENSRNOT00000017972 | STRG.22722.1.1 | STRG.22722.1 | 397.65 | 684.64 | 0.92 | 0.12 | 0.000457845 | 6.81E-01 | 0.014134119 | 9.71E-01 | 0.26 | 0.3 | 3.52 | 0.41 |
| Ap3b1 | ENSRNOG00000010624 | ENSRNOT00000008599 | STRG.7285.1 | STRG.7285.1 | 82.26 | 429.64 | 1.86 | 0.23 | 0.000463696 | 4.08E-01 | 0.01483458 | 9.14E-01 | 0.53 | 0.28 | 3.5 | 0.83 |
| Sirpa | ENSRNOG00000004763 | ENSRNOT00000009193 | STRG.14294.2 | STRG.14294.2 | 127.11 | 392.17 | 10.11 | 0.95 | 0.000464568 | 5.64E-02 | 0.014847706 | 4.87E-01 | 2.89 | 0.5 | 3.5 | 1.91 |
| Nsd1f3 | ENSRNOG00000011763 | ENSRNOT00000021103 | STRG.17741.2 | STRG.17741.2 | 89.03 | 112.97 | 1.3 | 0.64 | 0.000471784 | 2.39E-02 | 0.015003745 | 3.05E-01 | 0.37 | 0.28 | 3.5 | 2.26 |
| Zfp317 | ENSRNOG00000036684 | ENSRNOT00000078593 | STRG.30210.8 | STRG.30210.8 | 19.79 | 88 | 7.43 | 0.53 | 0.000513135 | 2.13E-01 | 0.01614316 | 7.99E-01 | 2.14 | 0.43 | 3.47 | 1.25 |
| Ptprr | ENSRNOG00000004483 | ENSRNOT00000006401 | STRG.27616.1 | STRG.27616.1 | 365.52 | 366.74 | 0.51 | 0.51 | 0.000528638 | 1.48E-03 | 0.016299416 | 4.25E-02 | 0.15 | 0.16 | 3.47 | 3.18 |
| Map1a | ENSRNOG00000014230 | ENSRNOT00000083065 | STRG.14044.1.1 | STRG.14044.1 | 120.3 | 2754.38 | 10.03 | 0.34 | 0.000526461 | 3.11E-01 | 0.016449725 | 8.74E-01 | 2.89 | 0.34 | 3.47 | 1.01 |
| Zfp707 | ENSRNOG00000057901 | ENSRNOT00000081186 | STRG.28862.10 | STRG.28862.10 | 111.56 | 178.85 | -0.83 | -0.42 | 0.000539342 | 6.37E-02 | 0.016754508 | 5.17E-01 | 0.24 | 0.22 | -3.46 | -1.65 |
| Slc44a1 | ENSRNOG00000035089 | ENSRNOT00000083983 | STRG.20579.2.1 | STRG.20579.2 | 772.72 | 1127.55 | -0.78 | -0.55 | 0.000588403 | 6.16E-03 | 0.018303222 | 1.26E-01 | 0.23 | 0.2 | -3.44 | -2.74 |
| Pkach | ENSRNOG00000049078 | ENSRNOT00000068739 | STRG.11630.1.1 | STRG.11630.1 | 123.64 | 295.87 | 10.07 | 1.24 | 0.000639328 | 1.60E-01 | 0.019239255 | 7.41E-01 | 2.95 | 0.88 | 3.41 | 1.4 |
| Srek1 | ENSRNOG00000012735 | ENSRNOT00000040533 | STRG.7519.2.1 | STRG.7519.2 | 443.45 | 387.65 | -0.78 | -0.21 | 0.000639673 | 5.99E-01 | 0.019239255 | 9.61E-01 | 0.23 | 0.41 | -3.41 | -0.53 |
| Prkce | ENSRNOG00000015603 | ENSRNOT00000020959 | STRG.23394.1 | STRG.23394.1 | 111.11 | 1985 | 9.92 | 0.99 | 0.000648752 | 2.48E-03 | 0.019421482 | 6.34E-02 | 2.91 | 0.33 | 3.41 | 3.03 |
| Fdxp3 | ENSRNOG00000044629 | ENSRNOT00000068531 | STRG.24817.1.2 | STRG.24817.1 | 120.32 | 2244.45 | -5.81 | -0.25 | 0.000651583 | 2.23E-01 | 0.019474188 | 8.08E-01 | 1.7 | 0.21 | -3.41 | -1.22 |
| Epb41l4b | ENSRNOG00000056550 | ENSRNOT00000090366 | STRG.20652.1 | STRG.20652.1 | 63.26 | 70.7 | -9.8 | -3.23 | 0.000667694 | 2.23E-01 | 0.019859121 | 8.08E-01 | 2.88 | 2.65 | -3.4 | -1.22 |
| Acn2 | ENSRNOG00000024128 | ENSRNOT00000038612 | STRG.29172.1.2 | STRG.29172.1 | 74.64 | 10501.67 | -10.03 | -0.07 | 0.000736359 | 6.28E-01 | 0.021542893 | 9.65E-01 | 2.97 | 0.14 | -3.38 | -0.48 |
| Ndufa4 | ENSRNOG00000011383 | ENSRNOT00000015237 | STRG.7784.1 | STRG.7784.1 | 1512.14 | 1848.59 | -0.48 | -0.32 | 0.000748414 | 7.43E-02 | 0.021779463 | 5.57E-01 | 0.14 | 0.18 | -3.37 | -1.78 |
| Kif21a | ENSRNOG00000014844 | ENSRNOT00000038946 | STRG.29470.2 | STRG.29470.2 | 103.88 | 235.68 | 9.82 | 1.21 | 0.000772041 | 7.46E-01 | 0.022382911 | 9.78E-01 | 2.92 | 3.73 | 3.36 | 0.32 |
| LOC688655 | ENSRNOG00000005811 | ENSRNOT00000007637 | STRG.23974.1 | STRG.23974.1 | 7036.27 | 7063.73 | -0.4 | -0.41 | 0.000798735 | 1.89E-03 | 0.023011577 | 1.15E-02 | 0.12 | 0.13 | -3.35 | -3.11 |
| Card1 | ENSRNOG00000007834 | ENSRNOT00000010720 | STRG.27721.2 | STRG.27721.2 | 105.49 | 2593.58 | 9.84 | 0.13 | 0.000800571 | 5.81E-01 | 0.023043827 | 9.58E-01 | 2.94 | 0.23 | 3.35 | 0.55 |
| Cxbs1 | ENSRNOG00000007456 | ENSRNOT00000013845 | STRG.19823.1.1 | STRG.19823.1 | 391.87 | 1866.02 | -0.84 | -0.07 | 0.000875523 | 7.56E-01 | 0.024823476 | 9.78E-01 | 0.25 | 0.22 | -3.33 | -0.31 |
| Copa2 | ENSRNOG00000008744 | ENSRNOT00000021287 | STRG.14140.1.1 | STRG.14140.1 | 1087.65 | 850.14 | -0.84 | 0.1 | 0.000893565 | 7.07E-01 | 0.025268159 | 9.73E-01 | 0.25 | 0.26 | -3.32 | 0.38 |
| Nxph | ENSRNOG00000004753 | ENSRNOT00000006329 | STRG.14715.1 | STRG.14715.1 | 108.21 | 1942.22 | 9.88 | 0.56 | 0.000904602 | 1.20E-01 | 0.025468256 | 6.73E-01 | 2.98 | 0.36 | 3.32 | 1.55 |
| Chst1a | ENSRNOG00000046479 | ENSRNOT00000071207 | STRG.33509.2 | STRG.33509.2 | 186.54 | 187.13 | 0.76 | 0.75 | 0.000969276 | 1.57E-03 | 0.026561535 | 4.46E-02 | 0.23 | 0.24 | 3.3 | 3.16 |
| Gnb2 | ENSRNOG00000000368 | ENSRNOT00000076234 | STRG.56903.4 | STRG.56903.4 | 543.88 | 545.88 | 0.65 | 0.64 | 0.000986949 | 1.69E-03 | 0.027261505 | 4.75E-02 | 0.2 | 0.2 | 3.29 | 3.14 |

# APPENDIX C

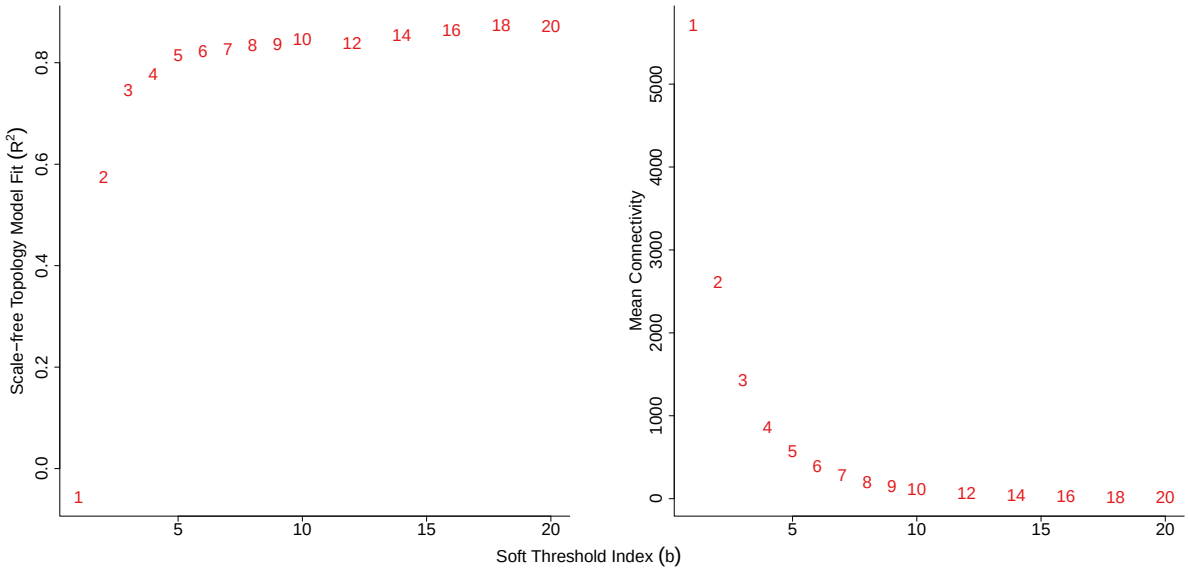
## CHAPTER IV SUPPLEMENTARY

### Supplementary Material

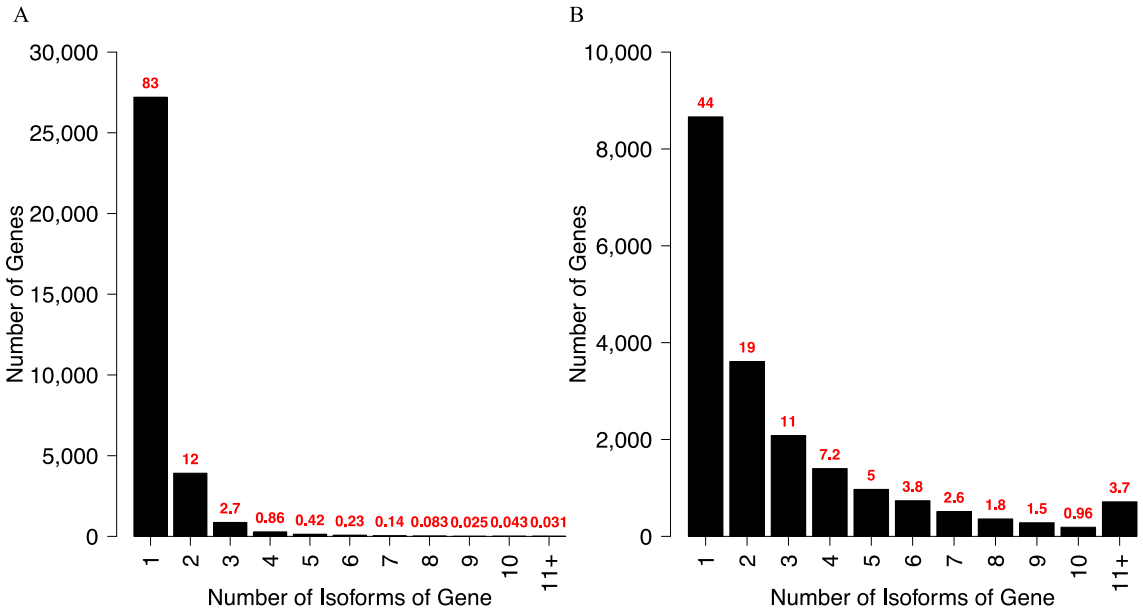
## Supplementary Figures



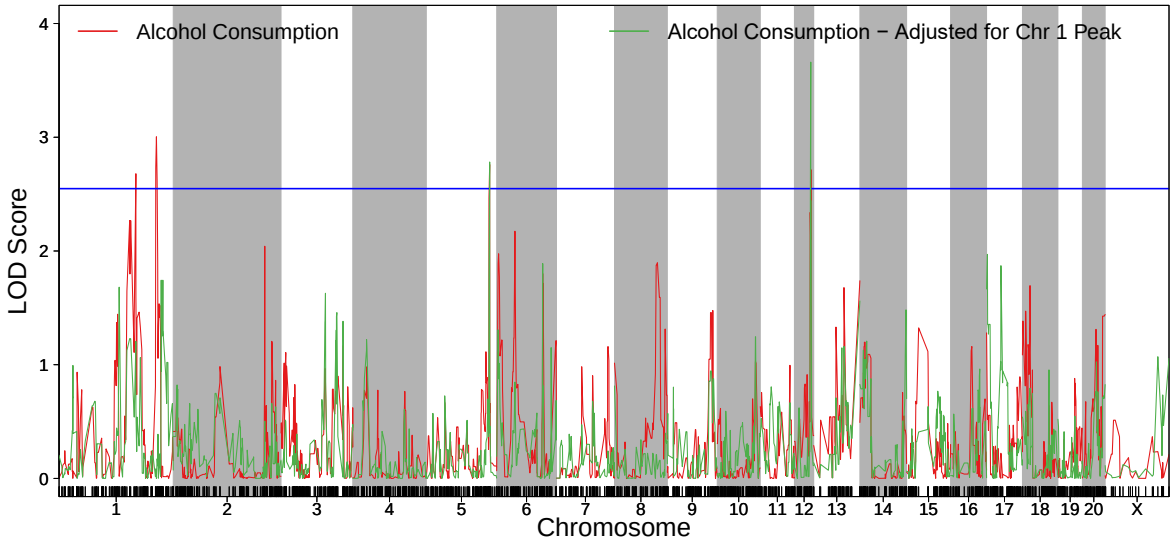

**Supplementary Figure 1. Outline of the transcriptome generation and quantitation steps.** **(A)** A general overview of the transcriptome generation steps used to generate the detection above background (DABG) transcriptome. Starting with the reference transcriptome, StringTiincorporated RNA sequencing (RNA-Seq) data to identify expressed transcripts. Aptardi likewise utilized RNA-Seq, as well as DNA sequence, to identify transcripts with different 3' termini than the input StringTie transcriptome. Finally, lowly expressed transcripts were removed by quantitating with RSEM and removing transcripts with zero counts in one third or more of samples or 200 bases or fewer in length. This was followed by re-quantitation and once again removal of transcripts with zero counts in one third or more of samples. Yellow boxes indicate transcriptomes, and blue boxes indicate data used during transcript assembly. **(B)** A detailed overview of the steps used to generate and quantitate the DABG transcriptome.




**Supplementary Figure 2. Network topology as a function of soft-thresholding power (β).** The influence of different index values on the goodness-of-fit to scale-free topology and topological features for the network were examined in **(A)** scale-free model fit index vs β and **(B)** mean connectivity vs β. Each point is labeled by its β index value in red. An index of seven was chosen for network construction.

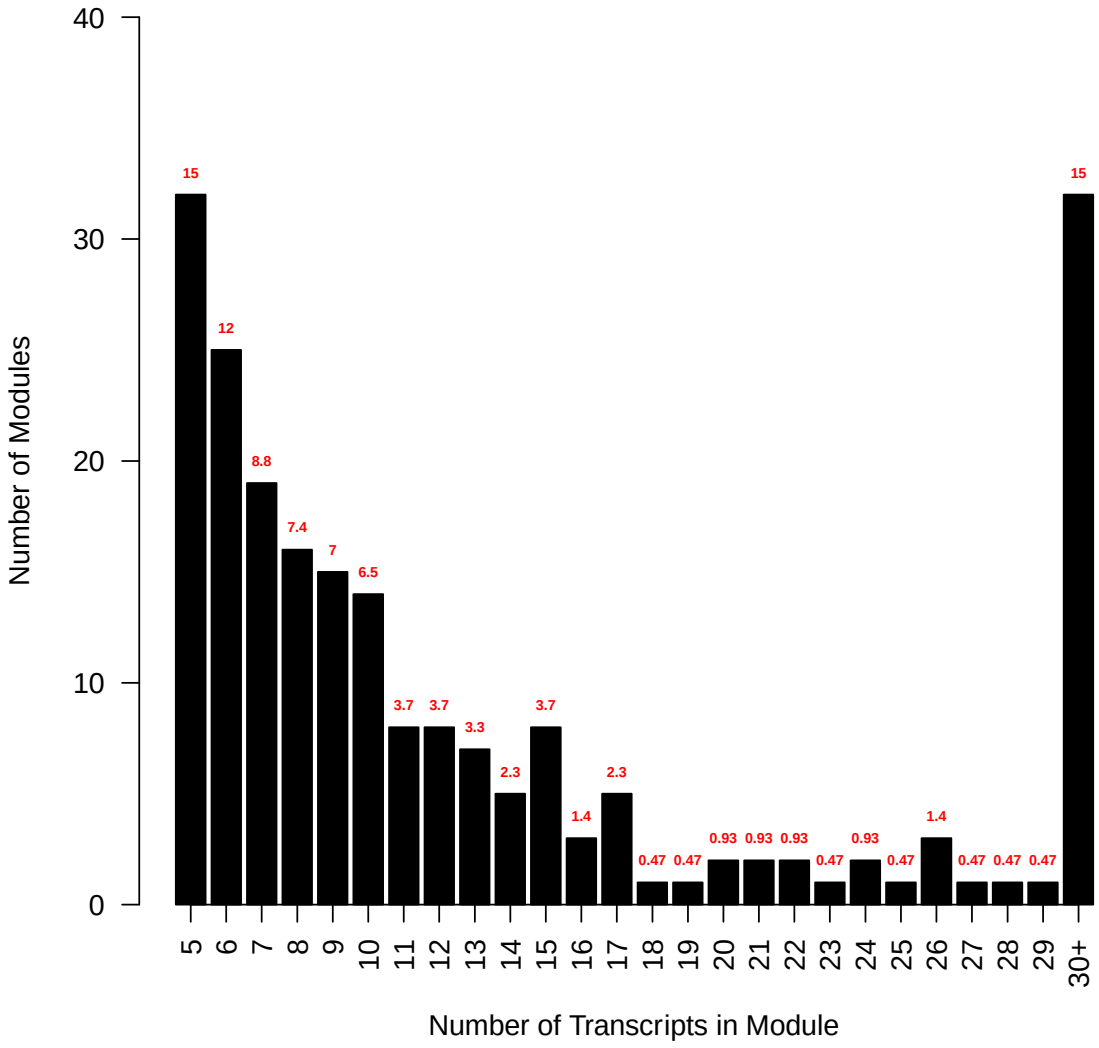



**Supplementary Figure 3. Number of isoforms for each gene in (A) the reference transcriptome and (B) the detection above background transcriptome.** Numbers above each bar indicate the percentage of total genes.

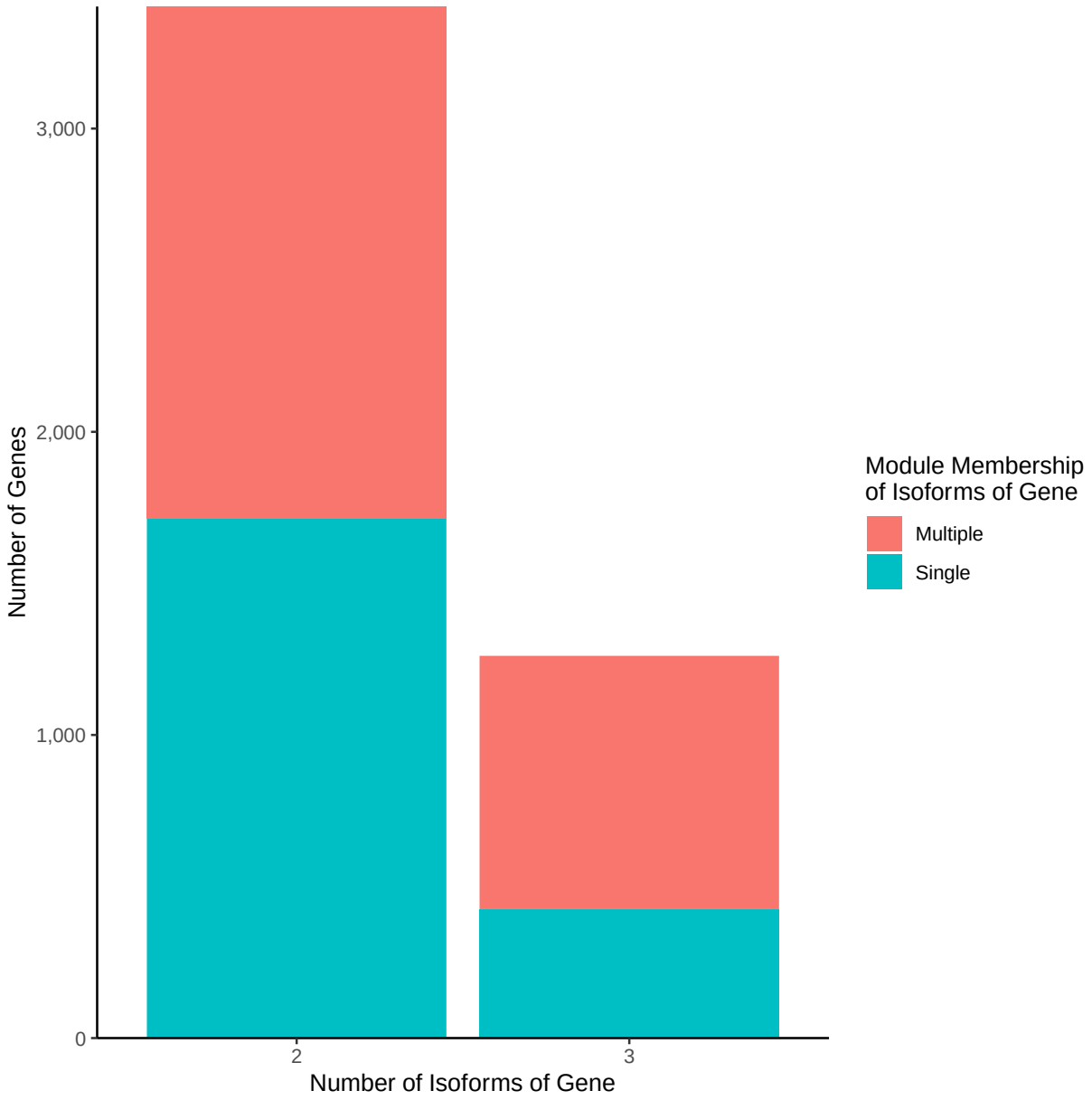



**Supplementary Figure 4. Voluntary alcohol consumption quantitative trait loci (QTL) adjusted for the maximum peak on chromosome 1.** The red plot represents the original QTL for alcohol consumption in the HXB/BXH recombinant inbred rat panel. The green plot represents the QTL scan for alcohol consumption after adjusting for the maximum peak chromosome 1 QTL (chr1:239 Mb). The blue line displays the logarithm of odds (LOD) suggestive threshold (p-value < 0.63) for genome-wide significance in the original analysis. The LOD scores below the suggestive threshold after adjusting for the maximum peak chromosome 1 QTL suggest that the two adjacent peaks on chromosome 1 display some degree of linkage disequilibrium and are better represented by a single QTL. The peaks on chromosome 5 and chromosome 12 remain suggestive after the adjustment, indicating that these two peaks contribute independently of chromosome 1 to the phenotype of voluntary alcohol consumption.

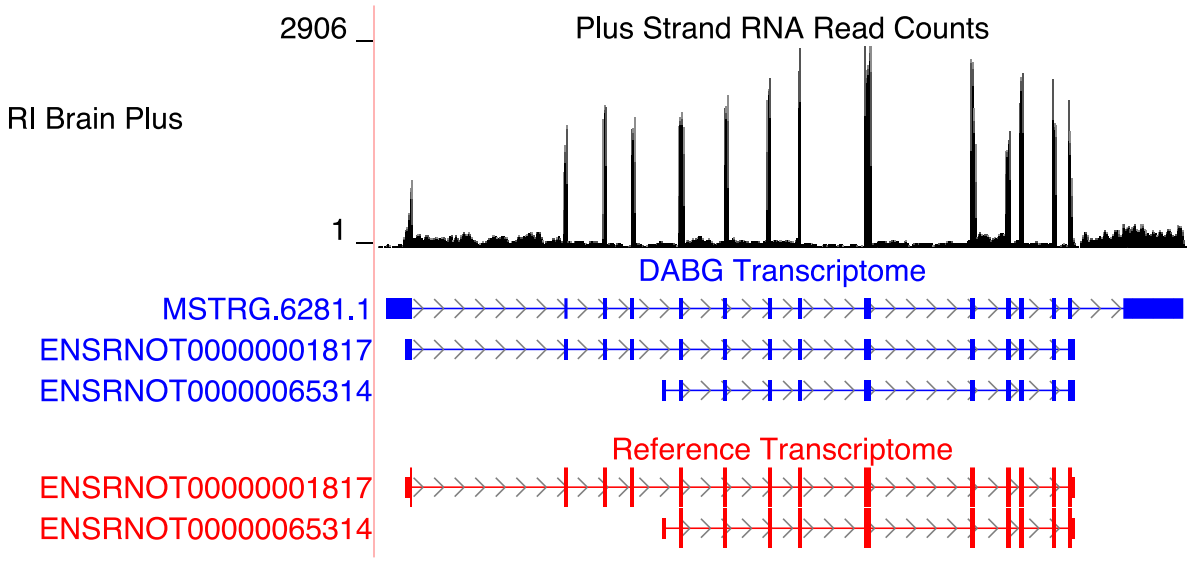

**Supplementary Figure 5. Outline of the steps used to identify candidate networks and transcripts.** Only transcripts with confident expression values were included by filtering for the top three expressed isoforms per gene using the mean transcript per million value across individual rat RNA sequencing samples for rats with alcohol consumption data only (63 samples, 21 strains). Next, only transcripts with high heritability (> 0.478) were included to focus on genetically influenced transcripts. Finally, transcripts that could be associated with a gene name were included for interpretability. (These steps are denoted by yellow boxes.) Using these transcripts, candidate coexpression modules were identified using weighted gene coexpression network analysis (WGCNA) and requiring 1) module eigengene association with voluntary alcohol consumption (Spearman's rank correlation coefficient p-value <= 0.01) and 2) a significant module eigengene QTL (genome-wide p-value < 0.01) overlap with voluntary alcohol consumption QTL (genome-wide p-value < 0.63) using 95% Bayesian credible intervals (green boxes). Individual candidate transcripts were identified by requiring 1) expression correlation with voluntary alcohol consumption (Spearman's rank correlation coefficient p-value <= 0.01) and 2) a significant expression QTL (genome-wide p-value < 0.01) overlap with voluntary alcohol consumption QTL (genome-wide p-value < 0.63) using 95% Bayesian credible intervals (orange boxes). Strain means of the normalized expression values were used as expression estimates for the transcripts in WGCNA/individual transcript associations, and strain mean voluntary alcohol consumption values were used for the association analyses. Heritability was estimated as the R-squared value from a one-way ANOVA using strain as the predictor (30 strains total) and transcript normalized expression estimates from individual rats as the response. Each strain possessed three expression estimates per transcript.




**Supplementary Figure 6. Distribution of module sizes built from weighted gene coexpression network analysis of the brain RNA expression data in the HXB/BXH recombinant inbred rat panel.** Module size is defined as the number of transcripts in the module. Modules with 30 or more transcripts were put into a single bin (30+).




**Supplementary Figure 7. Module memberships of transcripts derived from genes expressing more than one isoform and included in weighted gene coexpression network analysis.** The y-axis indicates the number of genes with more than one transcript (i.e. isoform), and the x-axis indicates the total number of isoforms of the gene. Blue bars indicate genes whose isoforms all belong to the same coexpression module, and red bars indicate genes whose isoforms belong to multiple modules.




**Supplementary Figure 8. Isoforms of the *Mapkapk5* gene.** Blue transcripts represent those identified in the detection above background (DABG) transcriptome, and the red transcripts represent the transcripts present in reference annotation. *ENSRNOT00000001817* and *ENSRNOT00000065314* are annotated in the reference transcriptome and retained in the DABG transcriptome, whereas *MSTRG.6281.1* represents a novel isoform identified here by StringTie. The RNA sequencing reads on the positive strand (black plot) represent a 10% randomly sampled subset from the HXB/BXH recombinant inbred rat panel RNA sequencing data in brain. This image was generated using the UCSC Genome Browser (http://genome.ucsc.edu).

## Supplementary Tables

### Supplementary Table 1. Overview of the strains with data available and the number of strains used in each analysis.

| Data type | Strains with data available | # strains |
|---|---|---|
| Phenotype: Voluntary alcohol consumption | BXH: 08, 10, 11, 12, 13, 6; HXB: 1, 10, 13, 15, 17, 18, 2, 20, 23, 25, 26, 27, 29, 3, 31, 4, 7 | 23 |
| RNA expression: Brain RNA expression | BXH: 10, 11, 12, 13, 2, 3, 5, 6, 8, 9; HXB: 1, 10, 13, 15, 17, 18, 2, 20, 21, 22, 23, 24, 25, 27, 29, 3, 31, 4, 5, 7 | 30 |
| Genotype: STAR consortium HXB/BXH recombinant inbred rat strain panel genotype data | BXH: 6, 10, 11, 12, 12a, 13, 2, 3, 5, 8, 9; HXB: 1, 2, 3, 4, 5, 7, 10, 13, 14, 15, 17, 18, 20, 21, 22, 23, 24, 25, 27, 29, 31 | 32 |
| | | |
| **Analysis** | **Data type(s) used** | **# strains used** |
| Voluntary alcohol consumption QTL | Phenotype and genotype | 21 |
| Correlation of individual transcript expression estimates with voluntary alcohol consumption | Phenotype and RNA expression | 21 |
| Transcript expression QTL analysis | RNA expression and genotype | 30 |
| WGCNA, brain coexpression module construction | RNA expression | 30 |
| Correlation of module eigengenes with voluntary alcohol consumption | Phenotype and RNA expression | 21 |
| Module eigengene QTL analysis | RNA expression and genotype | 30 |

The upper portion of the table includes the data types and strains possessing the data. The lower portion of the table includes the analysis type and datasets used for analysis. Only strains with data for each data type required in the analysis were included.

**Supplementary Table 2. The number of paired end reads from each HXB/BXH recombinant inbred rats strain RNA sequencing library after processing reads for quality.**

| RNA sequencing Library | # Paired End Reads |
|---|---|
| BXH12_1_brain_total_RNA_cDNA_GTCCGC | 148355651 |
| BXH12_2_brain_total_RNA_cDNA_CAGATC | 52079970 |
| HXB13_1_brain_total_RNA_cDNA_ATGTCA | 116868597 |
| HXB13_2_brain_total_RNA_cDNA_GTGAAA | 46829305 |
| HXB17_1_brain_total_RNA_cDNA_CCGTCC | 88744825 |
| HXB17_2_brain_total_RNA_cDNA_ATGTCA | 168985266 |
| HXB2_1_brain_total_RNA_cDNA_GTCCGC | 99182123 |
| HXB2_2_brain_total_RNA_cDNA_CTTGTA | 114362638 |
| HXB25_1_brain_total_RNA_cDNA_AGTTCC | 99324506 |
| HXB25_2_brain_total_RNA_cDNA_AGTCAA | 166249222 |
| HXB27_1_brain_total_RNA_cDNA_CGATGT | 105872835 |
| HXB27_2_brain_total_RNA_cDNA_AGTTCC | 124727782 |
| HXB7_1_brain_total_RNA_cDNA_ACAGTG | 115718346 |
| HXB7_2_brain_total_RNA_cDNA_AGTCAA | 162467043 |
| SHR_1_brain_total_RNA_cDNA_GCCAAT | 139282727 |
| SHR_2_brain_total_RNA_cDNA_TGACCA | 163139828 |
| BXH2_1brain_ATGTCA | 54395419 |
| HXB10_1brain_AGTCAA | 174354400 |
| HXB1_1brain_CGATGT | 144414084 |
| HXB15_1brain_AGTTCC | 126509012 |
| HXB18_1brain_AGTCAA | 139174364 |
| HXB20_1brain_AGTTCC | 116862005 |
| HXB21_1brain_CTTGTA | 129178807 |
| HXB22_1brain_GTGAAA | 124296887 |
| HXB23_1brain_GTCCGC | 154796799 |
| HXB24_1brain_GTCCGC | 66657290 |
| HXB29_1brain_CAGATC | 93213212 |
| HXB31_1brain_ATGTCA | 59230408 |
| HXB3_1brain_GCCAAT | 189101809 |
| HXB4_1brain_ACAGTG | 177983932 |
| HXB5_1brain_TGACCA | 199415470 |
| SHR_1brain_CCGTCC | 107459766 |
| BXH10_1brain_GCCAAT | 130991055 |
| BXH10_2brain_GTCCGC | 135753861 |
| BXH11_1brain_AGTCAA | 128875208 |
| BXH11_2brain_CTTGTA | 118956148 |
| BXH13_1brain_GTCCGC | 125318634 |
| BXH3_1brain_CGATGT | 133564814 |
| BXH3_2brain_CCGTCC | 90481113 |
| BXH5_1brain_ACAGTG | 134842078 |
| BXH6_1brain_TGACCA | 129481334 |
| BXH6_2brain_GTGAAA | 137408642 |
| BXH8_1brain_AGTTCC | 124708638 |
| BXH9_1brain_AGTCAA | 80099484 |
| BXH9_2brain_CAGATC | 108511730 |
| SHR_1brain_ATGTCA | 134043458 |
| SHR_1_brain_GTGAAA | 83012818 |
| SHR_3_brain_CGATGT | 95497387 |
| HXB10-2-brain-total-RNA_ATTACTCG | 48767922 |
| HXB10-3-brain-total-RNA_TAATGCGC | 65474217 |
| HXB13-3-brain-total-RNA_TCCGGAGA | 68178915 |
| HXB1-3-brain-total-RNA_GAGATTCC | 63557879 |
| HXB15-2-brain-total-RNA_CGCTCATT | 57679989 |
| HXB15-3-brain-total-RNA_CGGCTATG | 70562062 |
| HXB3-2-brain-total-RNA_CGGCTATG | 56804414 |
| HXB3-3-brain-total-RNA_ATTCAGAA | 62656125 |
| HXB4-2-brain-total-RNA_TCCGCGAA | 64497180 |
| HXB4-3-brain-total-RNA_GAATTCGT | 59027380 |
| HXB5-2-brain-total-RNA_TCTCGCGC | 35183682 |
| HXB5-3-brain-total-RNA_CTGAAGCT | 64659675 |
| HXB7-3-brain-total-RNA_AGCGATAG | 64578795 |
| SHR-1-brain-total-RNA_TCCGCGAA | 55423592 |
| BXH12_3-brain-total-RNA_S27 | 62692091 |
| BXH13_2-brain-total-RNA_S28 | 42327929 |
| BXH13_3-brain-total-RNA_S31 | 29413451 |
| HXB1-2-brain-total-RNA_S29 | 65476997 |
| HXB17_3-brain-total-RNA_S7 | 49976394 |
| HXB18_2-brain-total-RNA_S8 | 59529117 |
| HXB18_3-brain-total-RNA_S20 | 39953745 |
| HXB20_2-brain-total-RNA_S9 | 68334408 |
| HXB20_3-brain-total-RNA_S21 | 67600480 |
| HXB21_2-brain-total-RNA_S10 | 52687007 |
| HXB21_3-brain-total-RNA_S22 | 64305006 |
| HXB22_2-brain-total-RNA_S15 | 45596570 |
| HXB22_3-brain-total-RNA_S23 | 44230154 |
| HXB23_2-brain-total-RNA_S16 | 71551992 |
| HXB23_3-brain-total-RNA_S24 | 55804997 |
| HXB2-3-brain-total-RNA_S30 | 68638370 |
| HXB24_2-brain-total-RNA_S17 | 54265512 |
| HXB24_3-brain-total-RNA_S25 | 71170858 |
| HXB25_3-brain-total-RNA_S18 | 53458445 |
| HXB27_3-brain-total-RNA_S19 | 47836835 |
| SHR_1-brain-total-RNA_S26 | 84668795 |
| SHR_3-brain-total-RNA_S33 | 73238192 |
| BXH10-3-brain-total-RNA_CTGAAGCT | 77064133 |
| BXH11-3-brain-total-RNA_TAATGCGC | 79309805 |
| BXH2-2-brain-total-RNA_ATTACTCG | 64804424 |
| BXH3-3-brain-total-RNA_TCCGGAGA | 78382756 |
| BXH5-2-brain-total-RNA_CGCTCATT | 79239513 |
| BXH6-3-brain-total-RNA_GAGATTCC | 61524723 |
| BXH8-4-brain-total-RNA_ATTCAGAA | 71079477 |
| BXH9-3-brain-total-RNA_GAATTCGT | 80036840 |
| HXB29-2-brain-total-RNA_TCTCGCGC | 53467464 |
| HXB31-2-brain-total-RNA_AGCGATAG | 80119439 |
| BXH2-3-brain-total-RNA_S4 | 66768574 |
| BXH5-3-brain-total-RNA_S3 | 66915253 |
| BXH8-3-brain-total-RNA_S5 | 58113952 |
| HXB29-3-brain-total-RNA_S2 | 62702995 |
| HXB31-3-brain-total-RNA_S1 | 60567171 |
| SHR-1-brain-total-RNA_S6 | 64964188 |

**Supplementary Table 3. The genome alignment rate for each HXB/BXH recombinant inbred rat strain RNA sequencing library after processing for quality.**

| RNA sequencing Library | Strain Specific Genome Alignment |
|---|---|
| BXH12_1_brain_total_RNA_cDNA_GTCCGC | 98.28% |
| BXH12_2_brain_total_RNA_cDNA_CAGATC | 78.92% |
| HXB13_1_brain_total_RNA_cDNA_ATGTCA | 97.78% |
| HXB13_2_brain_total_RNA_cDNA_GTGAAA | 86.06% |
| HXB17_1_brain_total_RNA_cDNA_CCGTCC | 97.17% |
| HXB17_2_brain_total_RNA_cDNA_ATGTCA | 97.94% |
| HXB2_1_brain_total_RNA_cDNA_GTCCGC | 96.68% |
| HXB2_2_brain_total_RNA_cDNA_CTTGTA | 97.84% |
| HXB25_1_brain_total_RNA_cDNA_AGTTCC | 97.97% |
| HXB25_2_brain_total_RNA_cDNA_AGTCAA | 98.08% |
| HXB27_1_brain_total_RNA_cDNA_CGATGT | 97.34% |
| HXB27_2_brain_total_RNA_cDNA_AGTTCC | 96.87% |
| HXB7_1_brain_total_RNA_cDNA_ACAGTG | 97.94% |
| HXB7_2_brain_total_RNA_cDNA_AGTCAA | 98.22% |
| SHR_1_brain_total_RNA_cDNA_GCCAAT | 97.87% |
| SHR_2_brain_total_RNA_cDNA_TGACCA | 97.61% |
| BXH2_1brain_ATGTCA | 88.39% |
| HXB10_1brain_AGTCAA | 96.57% |
| HXB1_1brain_CGATGT | 97.19% |
| HXB15_1brain_AGTTCC | 96.12% |
| HXB18_1brain_AGTCAA | 97.17% |
| HXB20_1brain_AGTTCC | 97.34% |
| HXB21_1brain_CTTGTA | 97.17% |
| HXB22_1brain_GTGAAA | 97.87% |
| HXB23_1brain_GTCCGC | 97.27% |
| HXB24_1brain_GTCCGC | 94.76% |
| HXB29_1brain_CAGATC | 94.75% |
| HXB31_1brain_ATGTCA | 92.32% |
| HXB3_1brain_GCCAAT | 97.31% |
| HXB4_1brain_ACAGTG | 97.51% |
| HXB5_1brain_TGACCA | 97.42% |
| SHR_1brain_CCGTCC | 97.50% |
| BXH10_1brain_GCCAAT | 97.48% |
| BXH10_2brain_GTCCGC | 97.45% |
| BXH11_1brain_AGTCAA | 97.18% |
| BXH11_2brain_CTTGTA | 96.89% |
| BXH13_1brain_GTCCGC | 97.21% |
| BXH3_1brain_CGATGT | 97.03% |
| BXH3_2brain_CCGTCC | 94.90% |
| BXH5_1brain_ACAGTG | 97.28% |
| BXH6_1brain_TGACCA | 97.65% |
| BXH6_2brain_GTGAAA | 97.43% |
| BXH8_1brain_AGTTCC | 97.56% |
| BXH9_1brain_AGTCAA | 95.44% |
| BXH9_2brain_CAGATC | 97.16% |
| SHR_1brain_ATGTCA | 97.36% |
| SHR_1_brain_GTGAAA | 97.29% |
| SHR_3_brain_CGATGT | 97.15% |
| HXB10-2-brain-total-RNA_ATTACTCG | 96.38% |
| HXB10-3-brain-total-RNA_TAATGCGC | 96.32% |
| HXB13-3-brain-total-RNA_TCCGGAGA | 96.77% |
| HXB1-3-brain-total-RNA_GAGATTCC | 97.17% |
| HXB15-2-brain-total-RNA_CGCTCATT | 96.99% |
| HXB15-3-brain-total-RNA_CGGCTATG | 96.52% |
| HXB3-2-brain-total-RNA_CGGCTATG | 96.82% |
| HXB3-3-brain-total-RNA_ATTCAGAA | 96.98% |
| HXB4-2-brain-total-RNA_TCCGCGAA | 96.62% |
| HXB4-3-brain-total-RNA_GAATTCGT | 96.61% |
| HXB5-2-brain-total-RNA_TCTCGCGC | 97.04% |
| HXB5-3-brain-total-RNA_CTGAAGCT | 96.77% |
| HXB7-3-brain-total-RNA_AGCGATAG | 96.93% |
| SHR-1-brain-total-RNA_TCCGCGAA | 96.55% |
| BXH12_3-brain-total-RNA_S27 | 97.53% |
| BXH13_2-brain-total-RNA_S28 | 97.50% |
| BXH13_3-brain-total-RNA_S31 | 97.47% |
| HXB1-2-brain-total-RNA_S29 | 97.17% |
| HXB17_3-brain-total-RNA_S7 | 97.16% |
| HXB18_2-brain-total-RNA_S8 | 97.18% |
| HXB18_3-brain-total-RNA_S20 | 97.02% |
| HXB20_2-brain-total-RNA_S9 | 97.31% |
| HXB20_3-brain-total-RNA_S21 | 97.35% |
| HXB21_2-brain-total-RNA_S10 | 97.33% |
| HXB21_3-brain-total-RNA_S22 | 97.19% |
| HXB22_2-brain-total-RNA_S15 | 97.19% |
| HXB22_3-brain-total-RNA_S23 | 96.55% |
| HXB23_2-brain-total-RNA_S16 | 97.51% |
| HXB23_3-brain-total-RNA_S24 | 96.92% |
| HXB2-3-brain-total-RNA_S30 | 97.51% |
| HXB24_2-brain-total-RNA_S17 | 96.78% |
| HXB24_3-brain-total-RNA_S25 | 96.84% |
| HXB25_3-brain-total-RNA_S18 | 97.17% |
| HXB27_3-brain-total-RNA_S19 | 97.15% |
| SHR_1-brain-total-RNA_S26 | 96.92% |
| SHR_3-brain-total-RNA_S33 | 96.83% |
| BXH10-3-brain-total-RNA_CTGAAGCT | 97.28% |
| BXH11-3-brain-total-RNA_TAATGCGC | 97.35% |
| BXH2-2-brain-total-RNA_ATTACTCG | 97.50% |
| BXH3-3-brain-total-RNA_TCCGGAGA | 97.77% |
| BXH5-2-brain-total-RNA_CGCTCATT | 97.54% |
| BXH6-3-brain-total-RNA_GAGATTCC | 97.82% |
| BXH8-4-brain-total-RNA_ATTCAGAA | 97.28% |
| BXH9-3-brain-total-RNA_GAATTCGT | 97.78% |
| HXB29-2-brain-total-RNA_TCTCGCGC | 96.52% |
| HXB31-2-brain-total-RNA_AGCGATAG | 96.58% |
| BXH2-3-brain-total-RNA_S4 | 97.72% |
| BXH5-3-brain-total-RNA_S3 | 97.64% |
| BXH8-3-brain-total-RNA_S5 | 97.78% |
| HXB29-3-brain-total-RNA_S2 | 97.36% |
| HXB31-3-brain-total-RNA_S1 | 97.29% |
| SHR-1-brain-total-RNA_S6 | 97.33% |

Libraries were aligned to strain specific genomes using HISAT2.

**Supplementary Table 4. Unique 3’ termini identified by aptardi already annotated by a StringTie or reference transcript 3’ terminus (+/- 100 bases) before filtering (aptardi) and after filtering to generate the detection above background (DABG) transcriptome.**

| Dataset | # Unique 3' Aptardi Termini | % Unique Aptardi 3' Termini Corresponding to StringTie 3' Terminus (+/- 100 Bases) | % Unique Aptardi 3' Termini Corresponding to Reference 3' Terminus (+/- 100 Bases) | % Aptardi Transcripts Whose 3' Terminus Corresponds to StringTie or Reference 3' Terminus (+/- 100 Bases) |
|---|---|---|---|---|
| Aptardi | 34,003 | 13% | 12% | 21% |
| DABG | 14,388 | 11% | 8% | 16% |

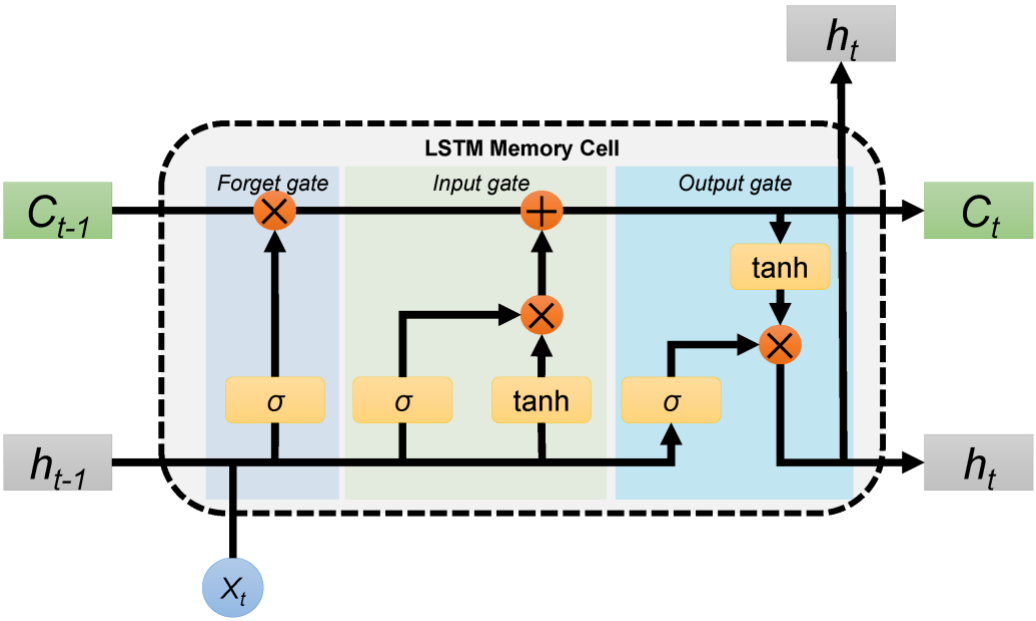
# APPENDIX D

# MACHINE LEARNING SUPPLEMENTARY

## Long short-term memory network memory cells

Key to memory cells in the LSTM is the presence of an internal state, or cell state (denoted C), in addition to the hidden state (denoted H). (Recall the hidden state represents information stored from the previous time steps.) The cell state possesses a self-connected recurrent edge that allows for error to traverse the memory cell without being changed due to a constant error carousel, which prevents vanishing/exploding gradients. In modern LSTM architectures, three gates control how information is used to update the cell state: the input gate, forget gate, and output gate (Figure 1.2). In the forget gate, also called the remember vector, input information from the previous hidden state and information from the current input are passed through a sigmoid activation function whose output is multiplied with the previous cell state. The forget gate was introduced after the original LSTM [395] and, intuitively, enables the LSTM to learn to reset itself at appropriate times, i.e. flushing of the contents of the cell state. In other words, the forget gate enables the reset of the constant error carousel, whereas in the original version of the LSTM the constant error carousel was maintained using a fixed weight of one for the cell state self-recurrent edge. The input gate likewise uses the previous hidden state and current input and passes this information through two parallel branches, one with an sigmoid activation and one with a tanh activation. The branch with sigmoid activation is akin to the standard way a recurrent neural net takes information from the input layer and from the hidden layer at the previous time step. Passing the information through the tanh function and multiplying the output of these branches together helps regulate the network. Intuitively, the input gate decides what information should be added to the cell state, i.e. long term memory. To

determine the current cell state, the previous cell state is first multiplied by the output of the forget gate. This value is then added to the output of the input gate to give the current cell state. Finally, the output gate, which determines the next hidden state, can be calculated using this newly determined current cell state. Namely, the output gate takes as input the previous cell state and information from the current input and passes it through a sigmoid activation function. Then the newly modified current cell state is passed through a tanh function and multiplied by the sigmoid output to yield the new hidden state used in subsequent time steps, along with the new current cell state.




**Supplementary Figure D.1. From [471].** Architecture of the long short-term memory cell and its operations. $C_t$ represents the cell state at time t, $C_{t-1}$ represents the cell state at time t-1, $h_t$ represents the hidden state at time t, $h_{t-1}$ represents the hidden state at time t01, and $x_t$ represents the input at time t.

**Gradient descent**

. Briefly, gradient descent minimizes the loss by taking derivative of the loss function and updating weights accordingly. An additional parameter, the learning rate, dictates the magnitude of weight adjustment at each update. Instead of updating the weights only after a complete

training set as in gradient descent, stochastic gradient descent constantly updates weights during training.

### Regularization techniques

Underfitting is typically addressed by additional features or more complex models. In the case of overfitting, regularization techniques such as Ridge, Lasso, elastic net, and dropout can be used where dropout is a technique specific to neural nets. Dropout is a regularization technique unique to artificial neural nets [472]. During training, nodes (and their connections) are randomly removed from the network to prevent co-adapting. Co-adapting occurs when nodes sharing connections learn together (i.e., specialize) by correcting each other's mistakes, which does not generalize to unseen data. Dropout effectively trains different network architectures in parallel which are then combined to produce the final model.

Both Ridge and Lasso penalize large coefficients (i.e. weights) by adding a penalty term to the loss. In Ridge regularization, also known as L2 regularization, the penalty term is the square of the coefficient, whereas in Lasso regularization, also known as L1 regularization, the penalty term is the absolute value of the coefficient. The penalty term is multiplied by a predetermined regularization parameter that determines the degree of regularization (i.e. a high regularization parameter confers greater regularization). An advantage of Lasso regularization compared to Ridge regularization is that Lasso enables weights to to converge to zero, which effectively acts as a feature selection method that produces sparse models. On the other hand, Lasso regularization cannot incorporate more features than the number of observations and, in the case of collinear variables, will randomly select one, which tends to lead to unstable models and loss of information. Ridge regularization does not suffer these traits. A third regularization

method, elastic net, uses both Ridge and Lasso regularization principles, which enables both feature selection while also inclusion of collinear variables.