the FEBS Journal

The sequenced rat brain transcriptome – its use in identifying networks predisposing alcohol consumption

Laura M. Saba[1], Stephen C. Flink[1], Lauren A. Vanderlinden[1], Yedy Israel[2], Lutske Tampier[2], Giancarlo Colombo[3], Kalervo Kiianmaa[4], Richard L. Bell[5], Morton P. Printz[6], Pamela Flodman[7], George Koob[8,*], Heather N. Richardson[8,†], Joseph Lombardo[9], Paula L. Hoffman[1,10] and Boris Tabakoff[1,10]

1 Department of Pharmaceutical Sciences, University of Colorado Denver, Aurora, CO, USA
2 Laboratory of Pharmacogenetics of Alcoholism, Molecular & Clinical Pharmacology Program, Institute of Biomedical Sciences, Faculty of Medicine, University of Chile, Santiago, Chile
3 Neuroscience Institute, National Research Council of Italy, Section of Cagliari, Monserrato, Italy
4 Department of Alcohol, Drugs and Addiction, National Institute for Health and Welfare, Helsinki, Finland
5 Department of Psychiatry, Institute of Psychiatric Research, Indiana University School of Medicine, Indianapolis, IN, USA
6 Department of Pharmacology, University of California San Diego, La Jolla, CA, USA
7 Department of Pediatrics, University of California, Irvine, Irvine, CA, USA
8 Committee on the Neurobiology of Addiction Disorders, The Scripps Research Institute, La Jolla, CA, USA
9 National Supercomputing Center for Energy and Environment, University of Nevada, Las Vegas, Nevada, USA
10 Department of Pharmacology, University of Colorado Denver, Aurora, CO, USA

Keywords
alcohol consumption; gene expression; HXB RI rat strains; neuroinflammation; rat selected lines

Correspondence
B. Tabakoff, Department of Pharmaceutical Sciences, Skaggs School of Pharmacy & Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus,
12850 E. Montview Blvd, Aurora,
CO 80045, USA
Fax: +1 303 724 6148
Tel: +1 303 724 3668
E-mail: boris.tabakoff@ucdenver.edu
Website: http://phenogen.ucdenver.edu

Present addresses:
*National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD, USA
†Department of Psychology, University of Massachusetts Amherst, Amherst, MA, USA

(Received 25 February 2015, revised 10 June 2015, accepted 23 June 2015)

doi:10.1111/febs.13358

A quantitative genetic approach, which involves correlation of transcriptional networks with the phenotype in a recombinant inbred (RI) population and in selectively bred lines of rats, and determination of coinciding quantitative trait loci for gene expression and the trait of interest, has been applied in the present study. In this analysis, a novel approach was used that combined DNA-Seq data, data from brain exon array analysis of HXB/BXH RI rat strains and six pairs of rat lines selectively bred for high and low alcohol preference, and RNA-Seq data (including rat brain transcriptome reconstruction) to quantify transcript expression levels, generate co-expression modules and identify biological functions that contribute to the predisposition of consuming varying amounts of alcohol. A gene co-expression module was identified in the RI rat strains that contained both annotated and unannotated transcripts expressed in the brain, and was associated with alcohol consumption in the RI panel. This module was found to be enriched with differentially expressed genes from the selected lines of rats. The candidate genes within the module and differentially expressed genes between high and low drinking selected lines were associated with glia (microglia and astrocytes) and could be categorized as being related to immune function, energy metabolism and calcium homeostasis, as well as glial–neuronal communication. The results of the present study show that there are multiple combinations of genetic factors that can produce the same phenotypic outcome. Although no single gene accounts for predisposition to a particular level of alcohol consumption in every animal model, coordinated differential expression of subsets of genes in the identified pathways produce similar phenotypic outcomes.

Database

The datasets supporting the results of the present study are available at http://phenogen.ucdenver.edu

Abbreviations

DABG, detection above background; FDR, false discovery rate; QTL, quantitative trait loci; RI, recombinant inbred; SNP, single nucleotide polymorphism; WGCNA, weighted gene co-expression network analysis.

Introduction

The rapid evolution of gene array technology from an expensive process with limited scope to an inexpensive, high throughput genome-wide interrogation of transcript levels has revolutionized genetic research. For example, the Affymetrix rat exon array (Affymetrix, Santa Clara, CA, USA) has over one million probe sets that interrogate the RNA expression levels of not only thousands of annotated protein-coding genes, but also thousands of predicted and not yet validated RNA transcripts. The ability to quantitatively measure the transcripts produced from an individual's DNA, generates a ubiquitous molecular endophenotype that has been shown to be of value in focusing the genetic analysis of complex quantitative traits to biological pathways important in the etiology of the trait of interest [1–3]. Although the technology related to gene arrays has vastly improved over the past 20 years, the technological drawbacks of using gene arrays such as the Affymetrix exon array platform include (a) different hybridization efficiencies across samples as a result of genomic variants [e.g. single nucleotide polymorphisms (SNPs) and indels] in the probed regions [4] and (b) annotation/interpretation issues related to different results from multiple probe sets targeting the same gene, or probe sets targeting more than one isoform of a gene.

To remedy these problems, we first utilized information from high throughput DNA sequencing on relevant samples to mask probes on the array that would be sensitive to differences in hybridization efficiency as a result of genetic variants within a probed region. We then used deep high throughput RNA sequencing information to identify known and novel transcripts expressed in a specific tissue (e.g. brain). With comprehensive information on the tissue-specific transcriptome, we evaluated and combined probe sets that provide information on splice variants of protein-coding genes, as well as annotated and unannotated noncoding transcripts expressed in the tissue, aiming to 'clean' the exon array data and improve the interpretation of expression estimates.

Once the use of our genetic and transcriptome information produced reliable and informative RNA expression levels from the exon array, we used this information to examine a complex behavioral trait (i.e. alcohol consumption). Alcohol consumption is considered to be the etiologic essential in the development of alcohol addiction [5–7], and levels of alcohol consumption by humans and other animals have been shown to have a strong genetic component [8,9]. In studies of concordance of alcohol consumption in monozygotic and dizygotic human twins, heritability for both the frequency and quantity of alcohol consumed varies between 0.4 and 0.7 [10,11]. The quantitative phenotype of alcohol consumption in both humans and rodents can be considered a polygenic trait [1,12–14], with several areas of the genome contributing to this phenotype.

Often with such polygenic, complex traits, the same genomic variant or the identical combination of genomic variants is not present in all individuals who manifest a particular phenotype. Instead, there are multiple variants or combinations of variants that produce the same diagnostic category. It is not a single genomic variant that is directly responsible for variation in a complex trait; instead, it is the effect of several, not always identical, genomic variants on the function of the biological pathway responsible for the phenotype that is the determining feature of genotype–phenotype relationships. One of the genetic tools for examining the plausibility of such claims is selective breeding. Selective breeding is a technique used to fix genetic elements that contribute to a trait of interest at the same time as hypothetically allowing for random recombination of other elements in the selected lines [15]. By conducting selective breeding under different selective pressures and/or with a different gene pool in the progenitors from which selection is initiated, one, in essence, can produce selection and fixation of different genes, which produce the same separation of phenotypes.

To our knowledge, there are currently six pairs of rat lines throughout the world, selected for high and low levels of alcohol consumption. Initial efforts to identify common differentially expressed genes in particular brain areas of various pairs of high drinking and low drinking lines have produced uninterpretable results [16,17] and it has been suggested previously by ourselves [18,19] and others [20] that one should consider a search for responsible networks rather than responsible genes.

But how does one identify relevant physiologic networks? Common ontology or cell type enrichment analyses may fall short when genes are under-annotated or even unannotated, such as for many of the noncoding transcripts identified in RNA-Seq datasets. An alternative is to incorporate another useful rodent model for examining complex traits [i.e. recombinant inbred (RI) strains]. The use of RI strains is a well-characterized and accepted technique for generating QTL and other quantitative genetic information [21]. RI panels allow not only for quantitative genetic analysis of behavioral phenotypes, such as alcohol consumption, but also RNA expression levels. In RI panels, the relationship between levels of expression of various genes has also been used for segregation of genes into networks (modules) by means of co-expression analysis, and this approach has been validated by studies demonstrating that modules of co-expressed genes are often strongly enriched in functional categories, or related to particular cell types [22,23]. A popular approach for deriving co-expression modules using gene expression data is weighted gene co-expression network analysis (WGCNA) [24].

In addition to using these co-expression modules to provide information about the physiologic function of genes differentially expressed among selected lines, they can be used to directly study the relationship between module expression patterns and alcohol consumption. To do this, the expression pattern of the whole module is summarized using a quantitative feature: its ‘eigengene’ (the first principal component of the gene expression matrix) [24]. The quantitative nature of the eigengene values allows for quantitative genetic analysis, including genetic correlations with alcohol consumption, and the use of quantitative trait loci (QTL) analyses to identify regions of the genome that control expression of the genes within the module. We have proposed that QTL overlap between a module eigengene and a phenotypic trait provides additional evidence showing that the functional characteristics or cell types represented by the genes included in the co-expression module play a role in the phenotype of interest, when genetic correlation between the eigengene and the trait has been established [25].

In the present study, the RNA expression estimates gathered with the ‘cleaned’ Affymetrix Rat Exon Arrays were combined with genotype and behavioral information in an extensive analysis that focused on the identification of a common functional pathway across both genetic models relevant to a predisposition for high or low alcohol consumption/preference. In the process, we generated a large volume of data on the transcriptional characteristics of the rat brain and mapped the expressed transcripts to strain-specific genomes of rats. All of the genomic and transcriptome information in its raw and analyzed forms is available on our website (http://phenogen.ucdenver.edu).

Results

Identification of gene/isoform probe set clusters

DNA and RNA sequencing

Of the approximately 1.7 billion read fragments (850 million paired-end reads) generated from the DNA of the two progenitor strains, 1.6 billion (96%) aligned with the rat reference genome. SNPs and small indels were identified for each strain separately with respect to the BN reference sequence (RGSC 5.0/rn5; http://genome.ucsc.edu). As expected, fewer SNPs and small indels (51 329 SNPs/66 470 small indels) were identified in the genome of the BN-Lx strain because it is a congenic of the BN reference strain [26]. In the SHR strain, 3 578 145 SNPs/1 089 050 small indels were identified compared to the reference BN genome. The SNPs and small indels of the sequenced genomes for BN-Lx and SHR strains are included in the genome browser available at http://phenogen.ucdenver.edu.

For the RNA-Seq data, over 1.6 billion read fragments (approximately 800 million paired-end reads) derived from both polyA+-selected RNA and ribosomal RNA-depleted total RNA were generated across the six brain samples (three BN-Lx rats and three SHR rats). Of those, more than 1.2 billion aligned with their respective strain-specific genomes. Combining the reconstructed transcriptomes from the total RNA and from the polyA+ RNA, and from both strains, resulted in 57 534 unique high confidence transcripts (35 511 unique genes). The characteristics of these transcripts and their overlap with current annotation are provided as an interactive graphic at http://phenogen.ucdenver.edu/PhenoGen/web/graphics/transcriptome.jsp.

Over 4.1 million probe sequences from the Affymetrix Rat Exon Array 1.0 ST were downloaded from the Affymetrix website (http://www.affymetrix.com). Of these, 3 664 621 (89%) aligned perfectly and uniquely with the reference BN rat genome and therefore were retained for further consideration. In addition, 108 563 (3%) of the retained probes were eliminated because they aligned with the rat genome in a region that harbored a SNP or small indel identified in the DNA-Seq data of the BN-Lx or SHR rats. The remaining ‘high integrity’ probes were summarized into 890 607 probe sets where at least three probes defined the probe set. When these probe sets were aligned with the brain transcriptome, we were able to create probe set clusters that represent 18 253 genes, as well as 19 023 probe set clusters for transcripts representing individual isoforms expressed in rat brain (Fig. S1).

Identification of candidate genes associated with a predisposition to alcohol preference/consumption

Selected lines meta-analysis

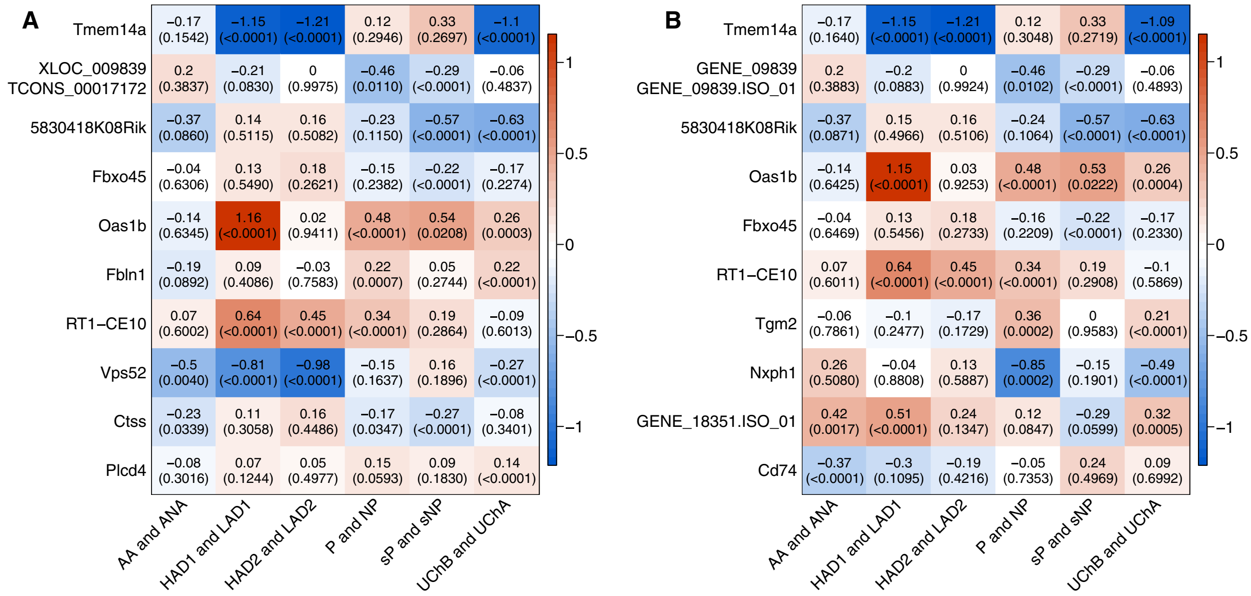
The differential expression meta-analysis of the selected lines was performed separately at the gene level and at the isoform-specific level. Of the 18 253 genes expressed in rat brain, according to the RNA-Seq data, and interrogated by the array, 16 074 genes were detected above background on the exon array in the selected lines [detection above background (DABG): $P < 0.0001$ in at least 5% of samples]. In addition, 123 genes were (a) differentially expressed [meta-analysis false discovery rate (FDR) < 0.05] and (b) showed a consistent direction of differential expression among individual line pairs that had statistical evidence ($P < 0.05$) for differential expression. The top ten differentially expressed genes based on the meta-analysis *P* values are shown in Fig. 1A. In the isoform-specific analysis, 14 594 transcripts were detected above background in the selected lines according to the array data and 95 were differentially expressed (meta-analysis FDR < 0.05) with the direction of differential expression consistent in pairs that had statistical evidence for differential expression. The top ten isoforms based on the meta-analysis *P* values are shown in Fig. 1B. Sixty-eight of the differentially expressed genes were represented in the list of differentially expressed isoforms. In other words, in these cases, the isoform expression contributed to the differential expression of the gene as a whole.

Alcohol consumption in HXB/BXH RI strains

Average daily alcohol consumption measures varied among strains in the RI panel (0.5–3.0 $g{\cdot}kg^{-1}$) (Fig. S2). Average daily alcohol consumption in this panel has a relatively high heritability (39%). The set of 7430 SNPs that differed between RI strains with alcohol consumption information and could be placed in the rn5 version of the rat genome represented a high-density map for this panel (average distance between SNPs = 0.37 Mb). After detailed quality control, this high-density map was reduced to 813 unique strain distribution patterns (i.e. haplotype blocks) across the 21 RI strains that had both genotype and alcohol consumption information. The bQTL analysis identified two peaks (Fig. 2) with suggestive genome-wide *P* values based on 1000 permutations (genome-wide *P* value threshold = 0.63; LOD = 2.39) [27].

WGCNA for RI strains

The brain RNA expression data gathered on 21 strains of the RI panel using the Affymetrix Rat Exon Array 1.0 ST were summarized into expression estimates for genes and isoforms. Separately, gene expression values



	AA and ANA	HAD1 and LAD1	HAD2 and LAD2	P and NP	sP and sNP	UChB and UChA
Tmem14a	−0.17 (0.1542)	−1.15 (<0.0001)	−1.21 (<0.0001)	0.12 (0.2946)	0.33 (0.2697)	−1.1 (<0.0001)
XLOC_009839 TCONS_00017172	0.2 (0.3837)	−0.21 (0.0830)	0 (0.9975)	−0.46 (0.0110)	−0.29 (<0.0001)	−0.06 (0.4837)
5830418K08Rik	−0.37 (0.0860)	0.14 (0.5115)	0.16 (0.5082)	−0.23 (0.1150)	−0.57 (<0.0001)	−0.63 (<0.0001)
Fbxo45	−0.04 (0.6306)	0.13 (0.5490)	0.18 (0.2621)	−0.15 (0.2382)	−0.22 (<0.0001)	−0.17 (0.2274)
Oas1b	−0.14 (0.6345)	1.16 (<0.0001)	0.02 (0.9411)	0.48 (<0.0001)	0.54 (0.0208)	0.26 (0.0003)
Fbln1	−0.19 (0.0892)	0.09 (0.4086)	−0.03 (0.7583)	0.22 (0.0007)	0.05 (0.2744)	0.22 (<0.0001)
RT1-CE10	0.07 (0.6002)	0.64 (<0.0001)	0.45 (<0.0001)	0.34 (<0.0001)	0.19 (0.2864)	−0.09 (0.6013)
Vps52	−0.5 (0.0040)	−0.81 (<0.0001)	−0.98 (<0.0001)	−0.15 (0.1637)	0.16 (0.1896)	−0.27 (<0.0001)
Ctss	−0.23 (0.0339)	0.11 (0.3058)	0.16 (0.4486)	−0.17 (0.0347)	−0.27 (<0.0001)	−0.08 (0.3401)
Plcd4	−0.08 (0.3016)	0.07 (0.1244)	0.05 (0.4977)	0.15 (0.0593)	0.09 (0.1830)	0.14 (<0.0001)

	AA and ANA	HAD1 and LAD1	HAD2 and LAD2	P and NP	sP and sNP	UChB and UChA
Tmem14a	−0.17 (0.1640)	−1.15 (<0.0001)	−1.21 (<0.0001)	0.12 (0.3048)	0.33 (0.2719)	−1.09 (<0.0001)
GENE_09839 GENE_09839.ISO_01	0.2 (0.3883)	−0.2 (0.0883)	0 (0.9924)	−0.46 (0.0102)	−0.29 (<0.0001)	−0.06 (0.4893)
5830418K08Rik	−0.37 (0.0871)	0.15 (0.4966)	0.16 (0.5106)	−0.24 (0.1064)	−0.57 (<0.0001)	−0.63 (<0.0001)
Oas1b	−0.14 (0.6425)	1.15 (<0.0001)	0.03 (0.9253)	0.48 (<0.0001)	0.53 (0.0222)	0.26 (0.0004)
Fbxo45	−0.04 (0.6469)	0.13 (0.5456)	0.18 (0.2733)	−0.16 (0.2209)	−0.22 (<0.0001)	−0.17 (0.2330)
RT1-CE10	0.07 (0.6011)	0.64 (<0.0001)	0.45 (<0.0001)	0.34 (<0.0001)	0.19 (0.2908)	−0.1 (0.5869)
Tgm2	−0.06 (0.7861)	−0.1 (0.2477)	−0.17 (0.1729)	0.36 (0.0002)	0 (0.9583)	0.21 (<0.0001)
Nxph1	0.26 (0.5080)	−0.04 (0.8808)	0.13 (0.5887)	−0.85 (0.0002)	−0.15 (0.1901)	−0.49 (<0.0001)
GENE_18351.ISO_01	0.42 (0.0017)	0.51 (<0.0001)	0.24 (0.1347)	0.12 (0.0847)	−0.29 (0.0599)	0.32 (0.0005)
Cd74	−0.37 (<0.0001)	−0.3 (0.1095)	−0.19 (0.4216)	−0.05 (0.7353)	0.24 (0.4969)	0.09 (0.6992)

Fig. 1. Genes/isoforms differentially expressed between high alcohol consuming and low alcohol consuming selected lines of rats. Genes/isoforms were ranked by *P* value from the meta-analysis including all six selected line pairs and the top ten genes (A) and isoforms (B) are included. Each row of the heatmap represents a gene/isoform and each column represents a selected line pair. The top line of each box is the $\log_2$ difference in expression (high consuming line – low consuming line). The bottom line is the *P* value for the difference in expression related to that particular pair. The colors of the boxes are based on the $\log_2$ difference in expression.

and isoform expression values that were detected above background in more than 5% of samples were subjected to WGCNA to identify co-expression modules. In the gene-level data, 364 modules were identified (median module size = 8 genes) (Fig. S3A) and, in the isoform-level data, 582 modules were identified (median module size = 7 isoforms) (Fig. S3B). The first principal component of each module (i.e. the eigengene) was calculated to represent the expression of genes/isoforms in the module across strains. In general, this eigengene captured a substantial portion of the variation among the genes and isoforms within each module (inter-quartile range: 61–70% in the gene-level analysis and 61–71% in the isoform-level analysis).

In the gene-level analysis, five modules were significantly associated with alcohol consumption using the combined *P* value (combined $P < 0.01$) (Table 1), which combined information on the correlation between the module eigengene and alcohol consumption in the RI panel and information on the enrichment of genes differentially expressed in the selected rat lines within the module. In the isoform-level analysis, five modules were significantly associated with

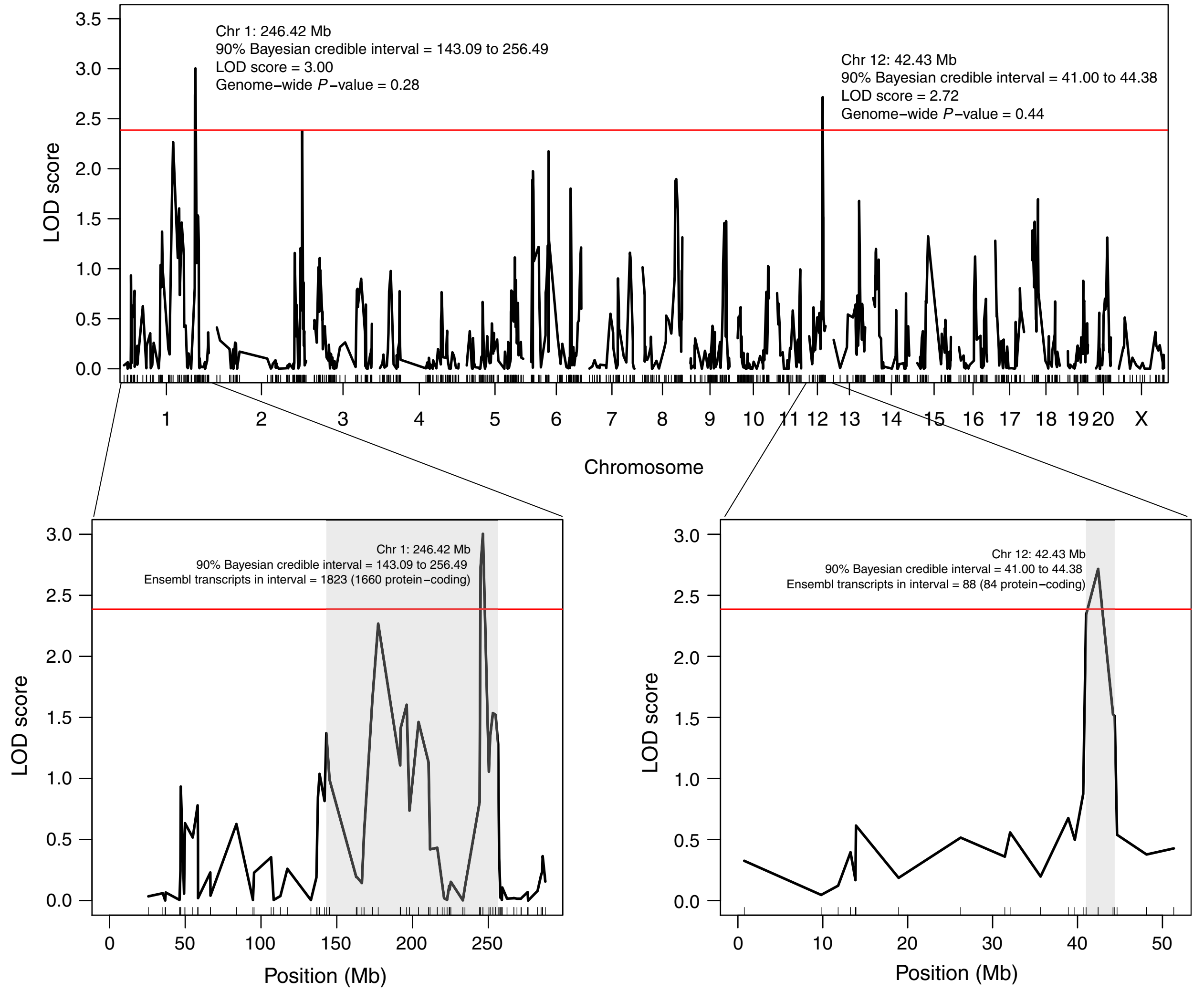


Fig. 2. LOD profile of voluntary alcohol consumption in the HXB/BXH recombinant inbred panel. Strain means were used in a marker regression to determine behavioral QTL for voluntary alcohol consumption in the two-bottle 24-h access paradigm. Two suggestive ($P < 0.63$) QTL are labeled with their location, credible interval, LOD score, and genome-wide *P* value. The red line represents the LOD threshold for a suggestive *P* value (2.39, genome-wide $P = 0.63$). The two insets are more detailed views of the two suggestive peaks. Their 90% Bayesian credible intervals are shaded in grey. The QTLs are labeled with their location, credible interval, and the number of Ensembl transcripts and Ensembl protein-coding transcripts physically located within each credible interval for a QTL.

alcohol consumption using the combined *P* value (combined $P < 0.01$) (Table 1).

We also examined the overlap between the module eigengene QTL and the QTL for alcohol consumption in the RI panel. Only one co-expression module in the gene-level analysis (indianred4) and one module in the isoform-level analysis (aquamarine1) passed this filter (Table 1 and Fig. S4). Many of the genes and isoforms are similar in these two modules. If a gene only had one splice variant expressed in the brain, the expression estimate at the gene level and at the isoform-specific level would be based on the same group of probe sets and would only deviate slightly as a result of normalization procedures. As a result of this overlap and because the eigengenes for the two modules were highly correlated (correlation coefficient = –0.73, $P = 0.0002$), we merged these two modules into one candidate co-expression module for visualization (Fig. 3).

In the combined module, a novel rat transcript (orthologous to A930024E05Rik in mouse and LOC101928346 in the human) was the most highly connected gene (i.e. the hub gene). The expression level of this novel transcript was highly heritable ($r^2 = 0.71$) in the RI panel, suggesting that its expression is under tight genetic control. Because this gene is not yet annotated in the rat genome, quantitative RT-PCR was used to verify the genomic structure of the transcript and the differential expression of the gene between the BN-Lx and SHR strains (detailed methods and results are provided in Doc. S1). In the transcriptome reconstruction, this gene consisted of two exons (Fig. 4). Based on a manual examination of the RNA-Seq reads and the correlation among probe sets from the Affymetrix Exon Array, there was evidence that an additional exon (from GENE 07345) could be included in this transcript (Fig. 4 and Doc. S1). The expression levels of three different fragments of the three-exon version of transcript were quantified by quantitative RT-PCR in the BN-Lx and SHR strains: (a) spanning exons 1 and 2; (b) spanning exons 2 and 3; and (c) spanning exons 1 and 3. Differential expression between strains (higher in SHR) was verified for all three fragments (all three $P < 0.001$). However, the expression levels of the three fragments within a strain were different. In both strains, the fragment spanning exons 1 and 2 had the highest expression level and the fragment spanning exons 1 and 3 had the lowest expression level (Doc. S1), indicating that multiple isoforms of this transcript may be expressed in rat brain. The clone produced from the primers that spanned exons 1 and 3 was sequenced and aligned with the genome. The first exon of the clone matched the exon that was not placed in GENE 07346 by the initial computational reconstruction. A large portion of this first exon is also found in human and mouse. The second (middle) exon of the clone was part of the computationally generated ‘gene’ and closely matched an exon from the orthologous mouse gene. The exon junction between the second and the third (final) exons matched precisely with the information from the reconstruction, but this exon was not present in the orthologous mouse gene (Fig. 4).

Characterization of common functional pathways among candidate genes

Although common ontology enrichment-based analyses can point one to general terms for annotating gene function, knowledge/literature-based analyses often uncover greater detail about functional pathways and potentially narrow or broaden views about the role that a particular transcript or pathway may play in the predisposition to a complex phenotype such as alcohol consumption. Knowledge/literature-based analyses are currently most effective when focusing on the aggregate of gene products, rather than on the individual isoform, and the knowledge/literature-based analysis is more easily applied to smaller sets of transcripts. The present study aimed to identify, with some confidence, the functional implications and interactions of the gene products that came to our attention through WGCNA (with the minimum module size reduced to capture smaller modules) and gene products that were brought to our attention through the meta-analysis of data derived from the six lines of rats selected for high and low consumption of ethanol.

In preparation for applying the knowledge/literature-based analyses, we combined results from our gene-level and isoform-level analyses and focused on gene products irrespective of isoform. In the selected lines meta-analysis, we reduced the list of 123 differentially expressed genes (FDR < 0.05) and 95 differentially expressed isoforms (FDR < 0.05) to the 10 genes and the 10 isoforms with the strongest statistical evidence of association with alcohol consumption (Fig. 1). Six gene products overlapped between the two lists (gene level and isoform level); therefore, the final set of candidates derived from the meta-analysis of the selected lines consisted of 14 unique gene products. In the WGCNA analysis, the gene-level and the isoform-level analyses had been combined to generate the candidate co-expression module (Fig. 3). This set of 18 candidate gene products was further reduced by requiring that their RNA expression levels also be

Table 1. Candidate modules from HXB/BXH WGCNA based on association with drinking in the RI panel and enrichment of differential expressed genes from selected lines analysis.

			Correlation with drinking in HXB/BXH		Enrichment for candidate genes from selected lines			Module eigengene QTL	
Module	Number of transcripts in module	Proportion of variance in module explained by eigengene	Correlation coefficient	*P* value	Enrichment *P* value	Candidate genes in module	Combined *P* value	Location [chromosome: Mb (95% confidence interval)]	Empirical genome-wide *P* value
Gene-level analysis									
indianred4	14	0.59	–0.59	0.005	0.0067	Tmem116;Oas1b	0.0004	chr12:41.0 (39.7–44.7)	0.033
brown	1743	0.51	–0.09	0.713	0.0001	intron of an est;no annotation;intron Gabrb3;intron of Rbfox1;intron Rbfox1; Fgf12;intron of Kalrn; intron GENE_09129.ISO_02; intron of Ptprg;intron Hs6st3; intron Fgf14;intron Ryr2;intron Ryr2;intron of Celf2;intron Gpr158;intron Kcnab1;no annotation;Ryr3;intron Tmem178b;intron of Ctnna2;intron of Magi1;intron of Xkr4;intron Nrxn1;intron of Dock4;covers Trim9;partial cover Mark3;Dync2h1;intron Ntm;intron of Arpp21	0.0006	chr3:177.6	0.312
plum2	18	0.61	–0.24	0.304	0.0006	Npas4;Btg2;Cyr61	0.0016	chr3:164.5	0.969
cyan	39	0.63	–0.11	0.646	0.0003	RT1-CE10;RT1-CE15;no annotation;Vps52	0.0021	chr20:8.0 (0.3–8.0)	0.008
brown2	14	0.62	0.06	0.792	0.0003	Sspn;Mettl20;Amn1	0.0022	chr4:244.0 (156.6–244.0)	0.095
Isoform-specific analysis									
maroon	18	0.66	–0.14	0.557	0.0004	RT1-A2;RT1-CE10;RT1-CE15	0.002	chr20:8.0 (0.3–8.0)	0.003
aquamarine1	8	0.61	0.33	0.145	0.0020	Tmem116;Oas1b	0.003	chr12:42.4 (40.7–44.7)	0.027
mediumorchid3	9	0.76	0.31	0.171	0.0025	no annotation;Scube2-ps1	0.004	chr1:177.3 (47.1–244.9)	0.006
grey60	31	0.62	0.60	0.004	0.1981	Idh1	0.007	chr2:278.0	0.966
lightpink3	14	0.58	0.54	0.011	0.0983	RGD1307461	0.009	chr8:121.7 (99.9–121.7)	0.006

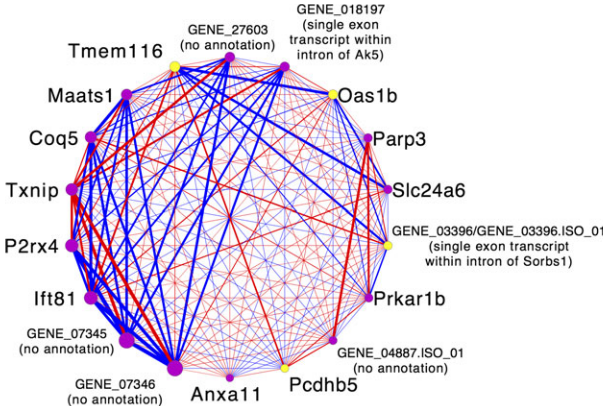


Fig. 3. Connectivity within the co-expression module associated with voluntary alcohol consumption. Each node represents a gene and/or an isoform from the two co-expression modules that were associated with alcohol consumption using a *P* value that combined information from the correlation of the eigengene with alcohol consumption and the enrichment of genes/isoforms within module differentially expressed in the rat lines selectively bred for high or low alcohol consumption. The size of the node is weighted based on its intramodular connectivity within the merged co-expression module. Nodes highlighted in yellow represent genes identified in both the gene-level analysis and the isoform-level analysis. The thickness of the line connecting two nodes (i.e. edge) is weighted based on the magnitude of the correlation coefficient between the two genes. Red edges represent a negative correlation and blue edges represent a positive correlation.

individually correlated ($P < 0.05$) with levels of alcohol consumption in the RI panel. This additional criterion that each transcript's independent correlation of expression levels with alcohol consumption levels eliminated eight gene products. These eight gene products were noted to have the lowest intramodular connectivity within the candidate co-expression module (Fig. 3). The 14 gene products from the selected lines study were combined with the remaining set of 10 gene products from the candidate co-expression module to create a candidate gene set for knowledge/literature-based analyses that contains 23 unique gene products (Table 2). Oas1b was part of the 14 gene products from the selected lines study and was part of the 10 gene products from the co-expression module. The characteristics of each of the candidate genes, including correlations among individual probe sets within the gene/isoform cluster, are described in Doc. S2. Of the 23 candidate genes, 12 had only one isoform in the transcriptome reconstruction. Three of the candidate genes had multiple isoforms, although only one of the isoforms was significantly associated with alcohol consumption (*Cd74*, *Tgm2* and *Nxph1*). In two of these three, the associated isoform was not the most highly expressed isoform of the gene according to the RNA-Seq data. These results may represent differences in isoform function, in that only one isoform is associated with alcohol consumption. Eight of the candidate genes had multiple isoforms, although they were only associated with alcohol consumption at the gene level. For most of these transcripts, the number of probe sets that could distinguish isoforms was limited, with some genes not having any probe sets that distinguished isoforms or probe sets that could only distinguish a minor isoform.

Using the GO database (http://www.geneontology.org/GO.database.shtml), the most significantly enriched biologic process category and the only GO term among the 23 candidate genes to reach statistical significance was immune response ($P = 0.03$). No GO terms from either the cellular composition category or the molecular function category were significantly enriched. When our gene list was subjected to a KEGG database analysis (http://www.genome.ad.jp/kegg), the top category was antigen processing and presentation ($P = 0.003$). The list of candidate genes was explored further by identifying enrichment using brain-derived lists compiled as part of the userListEnrichment function in the WGCNA R library [28]. Markers for three brain regions (hippocampus, frontal cortex and choroid plexus), four cell types (microglia, astrocytes, neurons and interneurons) and three intracellular domains (synaptic mitochondria, somatic mitochondria and cytoplasm) were over-represented within the candidate genes in Table 2 (Bonferroni adjusted $P < 0.05$). All of the categories above were utilized as 'concepts' defined by our candidate genes. We then proceeded to utilize the modification of the Formal Concept Analysis [29], which includes domain knowledge to explore the relationships among the 23 candidate gene products. The detailed results of this Concept Analysis are included in Doc. S3 and summaries of the results are provided in Fig. 5 and Table 3.

Discussion

The brain RNA-Seq data that we have gathered on the BN-Lx/Cub and SHR/Ola rats (and that we have made available on http://phenogen.ucdenver.edu) complement and significantly extend the recently published catalog of gene expression data from several organs of the Fisher 344 rat [30]. We have recently generated deep genome sequencing data for the F344 rat strain, which is currently available on our website and was published with the sequenced genomes of 40 other commonly used inbred strains of rats [31]. Given the RNA-Seq information provided previously

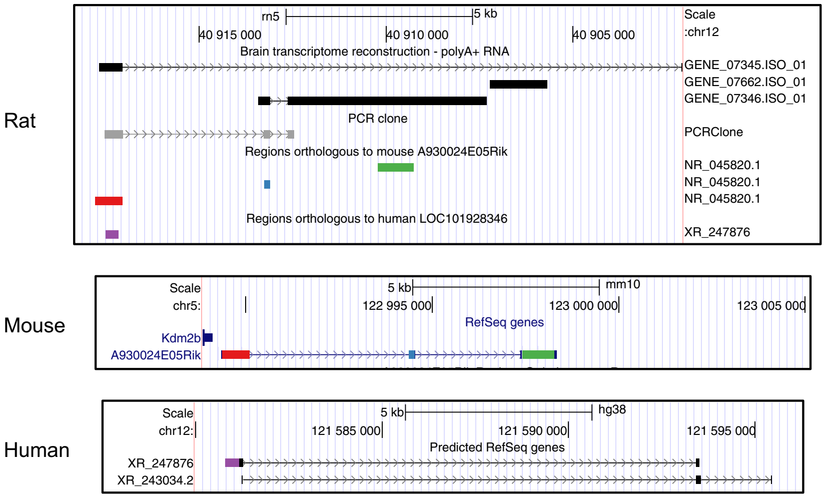


Fig. 4. Comparison of the transcriptome structure of novel rat transcript across mouse and human. The top box (Rat) is the genomic region, chr12:40,902,059-40,918,309, in the RGSC 5.0/rn5 version of the rat genome. In the rat, the novel transcript is transcribed from the negative strand. The numerical values of the coordinates have been reversed, 40 918 309 bp to 40 902 059 bp, so that the direction of transcription (left to right in the graphic) is consistent across species. In this box, the transcript structure of three transcripts derived from the transcriptome reconstruction using the polyA+-selected RNA is shown as the first series of tracks in black (e.g. GENE_07345.ISO_1). GENE_07346 is the hub gene for the co-expression module (Fig. 3). The second series (grey) in this box is the exon organization of GENE_07346 deduced from the PCR product sequence. The third series of tracks within this box indicate the genomic regions in the rat that are orthologous to the A930024E05Rik gene in mouse. The labels on the right are the relevant mouse RefSeq ncRNA ID. The final series of tracks in this box indicates the genomic region in the rat that is orthologous to LOC101928346 in humans. The label on the right is the relevant human RefSeq ncRNA ID. The second box (Mouse) is the genomic region, chr5:122,988,841-123,005,091, in the GRCm38/mm10 version of the mouse genome. The track within this box contains the A930024E05Rik gene as annotated in mouse. Regions that were identified as orthologous to the rat are colored with the same colors used in the Regions Orthologous to Mouse A930024E05Rik in the Rat box above. The third box (Human) is the genomic region, chr12:121,579,996-121,596,246, in the GRCh38/hg38 version of the human genome. The track within this box is the LOC101928346 lincRNA annotated in human with the relevant human RefSeq ncRNA IDs on the left. Regions that were identified as orthologous in the rat are colored with the same colors as in the Regions Orthologous to Human LOC101928346 in the Rat box above. It should be noted that the GENE_07346 and orthologous regions in the other two species are located between the Kdm2b and the Orai1 gene sequences. The figure was generated using the UCSC Genome Browser (http://genome.ucsc.edu).

[30], one can perform the same process that we have described in the present study for ‘cleaning’ the Affymetrix Exon Array data for use with the F344 strain, and for characterizing probes on the array that can identify specific expressed isoforms in the brain and other organs.

One notable extension to the published data is our inclusion of genome sequence and brain gene expression data across animals of different genetic backgrounds. By sequencing the genomes of rat strains that are the progenitors of the HXB/BXH RI panel of rats, we were able to characterize a large number of single base pair and indel polymorphisms in DNA between the parental strains. These polymorphisms are recombined in a diverse fashion across the RI panel and can be imputed into a strain-specific map for this RI panel [32] for quantitative trait analyses. The DNA-Seq data, combined with RNA-Seq data, served other purposes in the present study: (a) create a ‘mask’ to eliminate probes on hybridization-based gene expression arrays (Affymetrix Exon Arrays) that would produce erroneous results because of strain-specific differences in DNA/RNA sequence and (b) to aggregate and annotate probe sets based on the rat brain transcriptome derived from the RNA-Seq data. Through such a process, we generated a quantitative dataset from the Affymetrix Exon Arrays that was ‘polymorphism independent’ across the HXB/BXH RI panel, and made more definitive our search for pathways associated with a phenotype (levels of alcohol consumption). We have summarized the DNA and RNA sequencing data in the Results and the full data files are available at http://phenogen.ucdenver.edu. The data are also available in processed form through a genome browser on our website. The final versions of the masked Affymetrix Rat Exon Arrays are also available for download at http://phenogen.ucdenver.edu.

Table 2. Genes associated with a predisposition to variation in voluntary alcohol consumption.

Gene	Gene description	Analysis where gene was identified	Direction of association with drinking
5830418K08Rik	RIKEN cDNA 5830418K08 gene	Selected lines	Negative
Cd74	Cd74 molecule, major histocompatibility complex, class II invariant chain	Selected lines	Negative
Coq5	Coenzyme Q5 homolog, methyltransferase (*S. cerevisiae*)	Co-expression module	Negative
Ctss	Cathepsin S	Selected lines	Negative
Fbln1	Fibulin 1	Selected lines	Positive
Fbxo45	F-box protein 45	Selected lines	Negative
GENE_07345	Partial overlap with Orai1 and mouse A930024E05Rik	Co-expression module	Negative
GENE_07346	Homologous with mouse A930024E05Rik	Co-expression module	Negative
GENE_09839 GENE_09839.ISO_01	No annotation	Selected lines	Negative
GENE_18351.ISO_01	No annotation	Selected lines	Positive
GENE_27603	No annotation	Co-expression module	Negative
Ift81	Intraflagellar transport 81 homolog	Co-expression module	Negative
Maats1	MYCBP-associated, testis expressed 1	Co-expression module	Negative
Nxph1	Neurexophilin 1	Selected lines	Negative
Oas1b	2-5 Oligoadenylate synthetase 1B	Selected lines and co-expression module	Positive
P2rx4	Purinergic receptor P2X, ligand-gated ion channel 4	Co-expression module	Negative
Plcd4	Phospholipase C, delta 4	Selected lines	Positive
RT1-CE10	RT1-CE10 RT1 class I, locus CE10	Selected lines	Positive
Tgm2	Transglutaminase 2, C polypeptide	Selected lines	Positive
Tmem116	Transmembrane protein 116	Co-expression module	Positive
Tmem14a	Transmembrane protein 14A	Selected lines	Negative
Txnip	Thioredoxin interacting protein	Co-expression module	Positive
Vps52	Vacuolar protein sorting 52 homolog (*S. cerevisiae*)	Selected lines	Negative

By gathering high throughput RNA and DNA sequencing data on a few strains, we were able to vastly improve our large (over 300 samples) hybridization-based expression array data gathered prior to the explosion in efficiency of high throughput RNA sequencing. For example, previous studies have reported the detrimental effects of not accounting for SNPs within regions targeted by probes from hybridization arrays [4]. Such SNPs lead to false cis-eQTL and could result in co-expression patterns because of co-localization of genes rather than functional relationships. As seen from the results of the present study, the use of the reconstructed brain transcriptome from the RNA-Seq data on the progenitor strains identifies specific splice variants associated with alcohol consumption when more than one splice variant is detected in the brain. Also, the most highly connected transcript within the co-expression module associated with alcohol consumption is unannotated in the current rat transcriptome. In a typical microarray analysis, this probe set or probe set cluster may be eliminated because of its annotation ambiguity. However, within the context of the reconstructed transcriptome, we had an excellent starting point for identification of transcript structure through PCR, and we have been able to develop and refine a working hypothesis on its function within the context of brain. Using high throughput sequencing data, we were able to improve the accuracy of microarrays with respect to both RNA expression levels and the transcripts that they represent.

With our improved methods for estimating expression from microarray studies, we made use of two large expression data sets: (a) six pairs of rats selectively bred for alcohol preference and (b) a RI rat panel that displays a wide range of alcohol consumption values. Our hypothesis is that there are multiple genetic variants causing the same alcohol consumption phenotypes. For example, all of the studies that have compared genetic variants and differences in RNA

Table 3. Candidate transcripts in functional categories derived from formal concept analysis. (+) higher levels in high-drinking animals; (–) higher levels in low-drinking animals.

Generating and responding to immune signals	Glial/neuronal communication	Energy/redox/ calcium homeostasis
Ctss(–)	P2rx4(–)	
Txnip(+)	Vps52(–)	Plcδ4(+)
P2rx4(–)		P2rx4(–)
	Nxph1(–)	Maats1(–)
	Fbln1(+)	Coq5(–)
Cd74(–)	Ift81(–)	Tmem14a(–)
Oas1b(+)	Fbxo45(–)	Txnip(+)
Fbln1(+)		
Tgm2(+)	Tgm2(+)	Oas1b(+)
Nxph1(–)	Ctss(–)	
		Tgm2(+)
Summary: high drinking rats have lower innate immunity responsiveness	Summary: high drinking rats have lower purinergic transmission, lower GABA function, higher glutamate function	Summary: high drinking rats have lower glucose uptake and ATP production; lower cytosolic Ca^{2+}

expression levels among rat lines selected for alcohol preference have not identified a common gene across pairs of lines generated in different countries or even pairs of lines generated from a similar starting population at the same institution [16,17]. With differences in selective pressure and starting genetic pool, no gene or transcript was 'fixed' in the same manner in all six selectively bred pairs, although, in all cases, alcohol preference was altered as a result of breeding. However, there were many genes/transcripts that were fixed in more than one pair. If there was a single variant responsible for this phenotype, we would expect the same genetic variant to be identified in all selected line pairs. We therefore focused our attention on gathering information on strong candidate genes from the different rat models and then using these candidate genes in aggregate to infer a functional pathway involved in a predisposition to alcohol preference/ consumption.

In the RI panel, instead of observing two groups of rats with extreme alcohol consumption behaviors, alcohol consumption behaviors varied from high to low with many values in between. RI panels provide a useful tool for dissecting the effect of 'causal' variants on different genetic backgrounds and in combination with other causal variants with synergistic or opposing effects. For example, the two progenitor strains of the RI panel, SHR/Ola and BN-Lx/Cub, do not display extreme alcohol consumption behaviors. Instead, many RI strains consume less alcohol then either strain or consume more alcohol then either strain. This indicates that there are several causal variants for alcohol consumption in this panel and that the recombination of predisposing and protective variants determines the

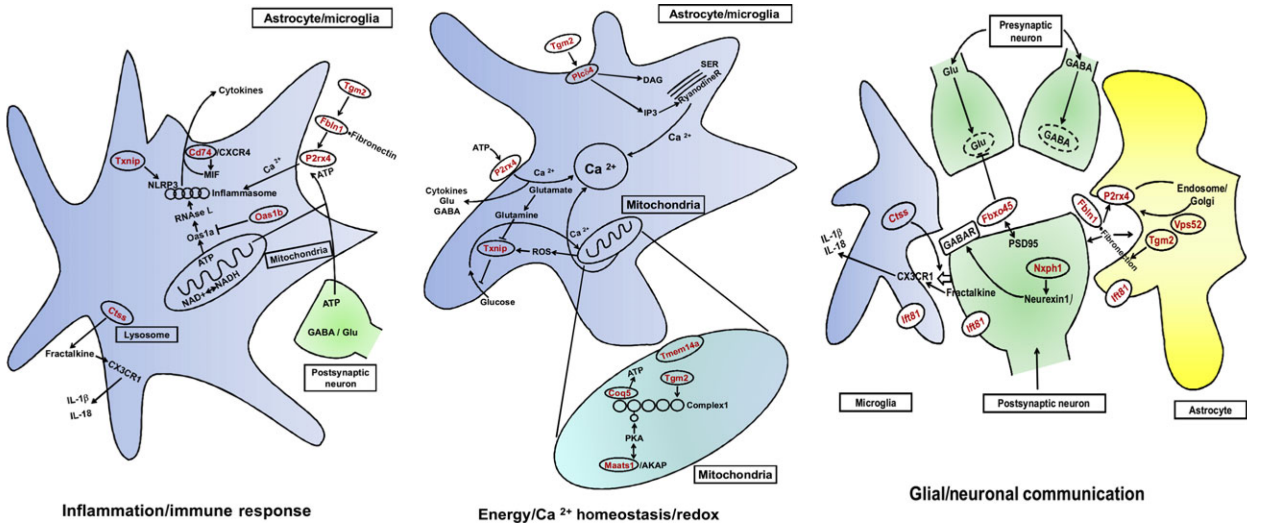


Fig. 5. Functional relationships among candidate genes for alcohol consumption. These cartoons illustrate the functions of and interactions among the annotated candidate genes for alcohol consumption that are described in more detail in the text, Doc. S3 and Table 3. Functions and interactions were derived from the Formal Concept Analysis and most candidate genes are expressed in glial cells (astrocytes and/or microglia). Each panel of the figure represents one of the functional categories listed in Table 3. Candidate genes are shown in red.

final phenotypic outcome. The RI panel contributed to the identification of a functional pathway related to alcohol consumption by first providing valuable information about the co-expression of genes/transcripts. Second, the panel provided information about expression QTLs. Third, the panel allowed for the direct examination of genetic correlation between RNA expression levels (via module eigengenes) and alcohol consumption. We used the co-expression information to identify modules of transcripts with similar expression patterns. Not only did this reduce the number of comparisons needed, but also it provided insight into the possible functional relationships between both well-annotated genes and under-annotated or unannotated RNA transcripts. We focused on co-expression modules that were enriched for genes/transcripts identified in the selected line study (i.e. different genes same pathway) and/or their expression, as measured through their module eigengene, was correlated with alcohol consumption.

We also included the criterion that the module expression QTL had to overlap a behavioral QTL for alcohol consumption. A number of studies have noted that variation in gene expression is a more prevalent mechanism underlying predisposition to complex (multifactorial) phenotypes [33,34] than genotypic differences that produce alterations in protein function. A clear mechanism for genetic control of the abundance of an RNA transcript is through polymorphisms in the regions coding for regulatory elements (e.g. sites for transcription factors and miRNAs, etc.). Such regulatory regions can control the expression of single transcripts and/or coordinately control the function of biological pathways [34].

In the HXB/BXH RI WGCNA, we changed the commonly used minimum threshold for module size from 30 to 5. We have used this adjusted threshold in other analyses, including the evaluation of modules for robustness, and have shown that even the smaller modules were ‘highly’ preserved in bootstrap samples [25]. However, to determine the sensitivity of the present study with respect to this adjusted threshold, we examined network results for gene-level data using the default parameters in the WGCNA analysis. This set of parameters identified 61 modules (compared to 364 using our original set of parameters). Using the same method for combining P values, we identified two modules with a combined $P < 0.05$. Although neither module had a significant module eigengene QTL, one module did have a ‘suggestive’ module eigengene QTL that overlapped a QTL for alcohol consumption. This module of 58 transcripts contained seven genes from Table 2. However, no gene ontology categories or KEGG pathways were enriched in this module ($P < 0.05$). Furthermore, neither of the associated modules in this network indicated both a correlation with drinking and an enrichment of differentially expressed genes from the selected lines. By contrast, our final candidate module in the gene-level data using the smaller minimum module size was both correlated with drinking in the RIs and enriched for differentially expressed genes from the selected lines (i.e. a more biologically robust result).

Our series of filters led us to one co-expression module generated from the combination of gene-level and isoform-specific analyses. This particular co-expression module also highlighted several of the benefits of using the high throughput sequencing data to inform the microarray analysis. First, the genes within the module were not all physically located near one another on the same chromosome. Therefore, we can conclude that SNPs within the probed regions are not artificially creating the observed co-expression patterns (i.e. not all genes have a cis-eQTL). Second, several of the transcripts were only included in the module because we could estimate the isoform-specific expression of those transcripts. More traditional ways of analyzing the data would have combined expression estimates from all isoforms of the gene, and the association with alcohol consumption would have been lost. Finally, several unannotated transcripts were contained in the module, including the most highly connected gene/transcript within the module. The transcriptome reconstruction provided additional information about the transcribed sequence of this gene and the inclusion of this transcript in this co-expression module gave us insight into possible functions of this transcript (Doc. S1).

Returning to the hypothesis that there are several ways to disrupt or alter the functional pathway responsible for variation in alcohol preference, the next step in our analysis was to identify a common function among the candidate genes identified in the different rat models. Accordingly, we needed to identify annotated genes with strong evidence for association with alcohol consumption. The goal was to start with our strongest evidence, with the knowledge that we are not trying to exhaustively identify every gene involved in the pathway, however we are trying to establish the identity of the involved pathway. We took the top genes from the selected lines meta-analysis and the genes from our candidate co-expression module that were individually correlated with alcohol consumption to begin our search of shared ontology and common annotated pathways. Using that information as a starting point, we did an in-depth literature review of the candidate genes (modified Formal Concept Analysis) to identify function, cellular location and interacting partners.

Overall, one can categorize the functions of the annotated proteins encoded by the 'candidate genes' into three major categories (with a number of gene products being included in more than one category, reflecting the significant cross-talk among these functional categories). The gene products are primarily associated with glia (microglia and astrocytes) and Table 3 lists the genes in each functional category: (a) Generating and Responding to Immune Signals; (b) Glial/Neuronal Communication; and (c) Energy, Redox and Calcium Homeostasis. Document S3 describes the functional characteristics of each gene product that align it with a particular category, and the relationships among these gene products are illustrated in Fig. 5. In summary, with respect to the Immune Signaling category, Txnip and P2X4 proteins influence the function of the NLRP3 inflammasome and modulate its caspase-dependent production and release of IL-1β and IL-18. Reactive oxygen species both activate *Txnip* transcription and promote the dissociation of Txnip from thioredoxin, allowing Txnip to perform functions such as activation of NLRP3. Cathepsin S, through proteolytic cleavage of fractalkine, which resides on neuronal membranes, produces a peptide that binds to and activates the CX3CR1 receptor located on both microglia and neurons, leading to release of interleukins [35]. The product of *Cd74* is part of a functional complex including the chemokine receptor, CXCR4. This complex can interact with the MIF protein produced in astrocytes and microglia to generate increases in the release of tumor necrosis factor-α, IL-8 and IL-1β.

Txnip not only participates in the innate immune response, but also is intimately involved in the energetics of microglia and astrocytes (Energy, Redox and Calcium Homeostasis) via its inhibition of glucose uptake [36]. Because glutamine, produced from glutamate by the glial glutamine synthetase, inhibits the transcription of *Txnip* and increases glucose uptake [36], Txnip can be considered as a key factor that modulates energy balance in glia. Also within the category of Energy, Redox and Calcium Homeostasis, Plcδ4 and the P2X4 receptor proteins are implicated in control of cytosolic calcium levels [37,38].

Mitochondrial ATP production and the resultant changes in NADH/NAD ratios are influenced by the products of other candidate genes in the category of Energy, Redox and Calcium Homeostasis (*Maats*, *Coq5* and *Tmem14a*). Transglutaminase 2 (*Tgm2*; expressed in neurons and glia) couples receptors to the activation of Plcδ, which is involved in inositol 1,4,5-trisphosphate and Ca^{2+} signaling. Tgm2 is also involved in maintaining the integrity of the mitochondrial respiratory complex 1 and 2 and maintaining ATP production [39]. The ATP produced by the mitochondria has numerous roles in the cell, and also functions as a transmitter in purinergic signaling (as a ligand for the P2X4 receptor on glia and neurons). Furthermore, ATP is a substrate for the oligoadenylate synthetase (Oas1a), which generates 2′-5′ oligoadenylates that are mandatory activators of RNAse L [40]. Oas1a activity is inhibited by Oas1b, which is the product of a candidate transcript, and recent evidence suggests that RNase L activation is an important component of the innate immune response [41]. Therefore, *Oas1b* can also be included in the Immune Response Category, as can the proteins that affect intracellular Ca^{2+} levels, because Ca^{2+} can activate the NLRP3 inflammasome [42].

With regard to Glial/Neuronal Communication, the interaction of cathepsin S and fractalkine was noted earlier. Purinergic receptor signaling is again evident in this category through the redundant presence of *P2rx4*. This is complemented by the presence of *Vps52*. The product of *Vps52* is a component of the endosome/Golgi/lysosome receptor recycling system that is involved in the rapid recycling of the P2X4 receptor occurring in neurons and glia. *Fbln1* codes for fibrulin, a small extracellular matrix protein, which binds to fibronectin. The fibrulin/fibronectin complex on the surface of glial cells (particularly microglia) promotes microglial activation, including increased transcription of *P2rx4* and increased delivery of this receptor to the cell surface [43]. The protein product of *Tgm2* also promotes the interaction of fibronectin with other proteins [44].

Neurexophylin 1 [45], a candidate in the Glial/Neuronal communication category, is present in neurons, and is processed to neurexin1α, which promotes the development of GABAergic synapses [46]. Other proteins generated by transcripts in the Glial/Neuronal Communication category include the ubiquitin ligase scaffolding protein Fbxo45, as well as Ift81. Fbxo45 is linked to glutamatergic transmission through its interactions with the cytokine-inducible form of nitric oxide synthetase, influencing glutamate release in neurons and astrocytes [47]. Fbxo45 also plays a direct role in inhibiting glutamatergic vesicle fusion with synaptic membranes and glutamate release [48]. Ift81 is a critical component of cilium formation in astrocytes and neurons [49]. This protein is affected by cytosolic Ca^{2+} levels [50], and the cilium is positioned to sense physical and biochemical extracellular signals, such as nutrients, and, in certain instances, modulate consummatory behavior [51].

The summary above indicates not only that several of the candidate gene products function within more than one category, but also that several of the candidate gene products can affect the same outcome by different mechanisms (e.g. Txnip and Cathepsin S both modulate the release of IL-1β from glia) and several of the candidate gene products impact steady-state cytosolic calcium concentrations and calcium responsive reactions. These observations reinforce the fact that, in different rat strains or lines, one can find differential expression of unlike genes, which, however, generate a similar neurobiological and behavioral outcome.

Table 3 categorizes the differentially expressed transcripts that may predispose animals to high drinking. If the characteristics of the pathways in which the identified gene products participate drive alcohol consumption, then alcohol may in some way interact with these pathways. When one considers how alcohol can interact with glial/neuronal communication, immune system function, and brain energy/redox and calcium dynamics, one has to carefully dissociate studies that measure the pathologic consequences of high levels of chronic ethanol intake from the impact of the ‘normal’ range of alcohol consumption on the brain networks that we have identified. Although the direct effects of ethanol on several systems identified by our studies have been examined (e.g. effects on the NLRP3 inflammasome complex) [52], the reported effects occurred under conditions where ethanol levels were much higher than those found in rats voluntarily consuming ethanol.

On the other hand, acetate, the metabolite of ethanol, may be a particularly important factor that affects the systems identified by our analysis. Acetate is formed in the liver from ingested ethanol released into the circulation and is found in significant quantities in the brain of rats and humans, even after low levels of ethanol exposure [53,54]. Acetate is converted to acetyl-CoA and metabolized via the TCA cycle primarily by astrocytes in the brain [55]. Acetate metabolism through the TCA cycle contributes to the synthesis of GABA, neurotransmitters glutamate/glutamine, ATP and lactate in astrocytes [56], which can all be released to modulate neuronal excitability and metabolism. Given the higher levels of *Txnip* expression in the high ethanol consuming rats, the Txnip could diminish glucose uptake into astrocytes, and the acetate derived from ethanol could ‘rescue’ astrocytic metabolism [57] and enhance the production of both GABA and glutamine/glutamate, as well as ATP, all of which play important roles in the genetic predisposition for variation in alcohol consumption [19].

Acetate can also affect the link between cellular (and particularly glial) energy metabolism, redox state and calcium homeostasis. Acetate can promote calcium release from mitochondria into cytosol [58] and reduction of cytosolic calcium requires energy in the form of ATP. The inherent differences in expression of transcripts related to energy metabolism and calcium homeostasis in the high versus low-drinking animals, in turn, interact with a myriad of effectors that influence brain function.

With regard to neuroimmune systems, alcohol drinking behavior may join a number of cognitive disorders (Alzheimer’s dementia, schizophrenia) and mood disorders (major depressive disorder, generalized anxiety disorder) that have been related to (mal)-function of neuroimmune systems [59,60]. The role of neuroinflammation and the immune system in alcohol consumption has been a focus of recent research [61–63]. Blednov *et al.* [64] found that the administration of lipopolysaccharide to mice normally consuming high levels of alcohol resulted in a further increase in alcohol consumption, although lipopolysaccharide did not affect alcohol consumption in a strain with low levels of alcohol consumption. Again, acetate becomes a factor when considering these results. Soliman *et al.* [65] have demonstrated that acetate can ameliorate lipopolysaccharide-induced astrocyte activation and cytokine release. Is alcohol drinking increased to reduce inflammation, or is a lower activity of the innate immune system promoting drinking? Our data suggest that, within the ‘normal’ range of function, the innate lower function of the immune mechanisms related to cytokine release, the MIF•CD74/CXCR2/CXCR4 signaling system and the cathepsin S/fractalkine/CX3CR1 system, may well diminish the levels of alcohol intake in a free choice situation.

It should be stressed that the present studies aimed to examine the brain transcriptional landscape to generate information that is predictive of levels of free choice ethanol consumption and not the result of alcohol consumption. Whether the same transcriptional networks are important in the escalation of ethanol consumption once alcohol intake has been initiated remains to be examined. Our previous work [66] does, however, indicate that the same bQTL, along with others, can be identified when examining changes in drinking by the HXB/BXH RI panel after 15 weeks of ethanol consumption and, interestingly, chronic ethanol consumption increases acetate production and brain acetate uptake in both rats and humans [57,67]. With respect to the translational relevance of the present study, three of the candidate transcripts that we identified (TXNIP, OAS1, PLCD4) were differentially expressed in postmortem hippocampal tissue of alcoholics compared to controls [68]. Additionally, a transcript identified as LOC101928346 with sequence homology and syntenic location similar to that of our module hubgene has been identified in humans (NCBI Reference Sequence: XR_247876). In work conducted on the post-mortem brain, the question of whether differentially expressed transcripts are involved in the predisposition (risk) for high levels of ethanol consumption, or are a result of alcohol consumption, remains unresolved. The results of the present study provide evidence that these transcripts and the pathways with which they are associated may mediate the predisposition (risk) for variation in alcohol consumption in animals including humans (i.e. are inherently expressed at different levels, rather than being altered in their abundance by chronic consumption of ethanol).

Materials and methods

Overview

The main goal of the present study was to identify functional pathways related to a predisposition to alcohol preference/consumption. To reach this goal, the analysis was split into three major steps: (a) identification of high integrity gene and isoform probe set clusters (Affymetrix Rat Exon 1.0 ST Array) based on the rat brain transcriptome; (b) identification of candidate genes associated with a predisposition to alcohol preference/consumption in RI strains and selected lines; and (b) characterization of common functional pathways among candidate genes (for a detailed work flow, see Fig. S1).

Identification of gene/isoform probe set clusters

To generate high integrity probe set clusters that were specific to genes and individual isoforms expressed in the brain, we generated high throughput sequencing data on both DNA and brain RNA in two common inbred rat strains (SHR/OlaIpcvPrin and BN-Lx/CubPrin rats; hereafter called SHR/Ola and BN-Lx/Cub rats) that not only represent genetic extremes among laboratory rats [69], but also represent the two progenitor strains of the HXB/BXH recombinant inbred panel [32] utilized in our alcohol consumption studies. The DNA sequence information provides guidance for the elimination of individual probes whose hybridization efficiency is compromised by SNPs or small insertions or deletions in our samples. The RNA sequence information provides guidance for construction of probe set clusters that represent genes expressed in rat brain and probe set clusters that estimate the expression of individual isoforms in rat brain.

Identification of candidate genes associated with a predisposition to alcohol preference/consumption

With the newly defined gene and isoform probe set clusters for the Affymetrix Rat Exon 1.0 ST array, we estimated RNA expression levels in two rat populations: the HXB/BXH RI panel and the six pairs of selectively bred rat lines. We used a meta-analysis approach to identify genes/isoforms differentially expressed among high and low alcohol consuming selected lines. We utilized WGCNA [24] to identify co-expression modules using gene expression data from the RI strains. To identify modules associated with alcohol consumption/preference, we relied on the convergence of evidence from (a) enrichment of genes/isoforms differentially expressed in the selected lines; (b) genetic correlation of the module eigengene with alcohol consumption in the HXB/BXH panel; and (c) overlap of the QTL for the module eigengene with a QTL for the alcohol consumption behavior measured in the HXB/BXH RI panel. This required several individual analyses and, in many of these analyses, we used 'liberal' thresholds for statistical significance (see below). We argue that the strength of the entire collection of data and the combined analyses is that data from several sources are used and convergence of evidence (even marginal evidence) instills confidence in the results.

Characterization of common functional pathways among candidate genes

The genes/isoforms with the most statistical evidence for association (i.e. lowest *P* values) with alcohol consumption in the selected lines were combined with genes/isoforms from the candidate co-expression module in the RI panel to form a list of candidate genes that represent the shared functional pathway responsible for predisposition to alcohol preference/consumption in rats. From this list of candidate genes, we utilized the modification of the Formal Concept Analysis [29], which includes domain knowledge (PubMed-derived information) to explore the relationships among the candidate gene products (BT acted as the 'domain expert'). To initiate this analysis, we first identified 'concepts' through functional and cell type enrichment analyses using the GO database (http://www.geneontology.org/GO.database.shtml), the KEGG database (http://www.genome.ad.jp/kegg/) and brain-derived lists compiled as part of the userListEnrichment function in the WGCNA R library [28]. The brain-derived lists include markers for brain region-specific expression, cell type-specific expression and expression specific to an intracellular domain.

Detailed methods for identification of gene/isoform probe set clusters

DNA-Seq

Genomic DNA was extracted from 25 mg of homogenized brain tissue from males of the progenitor strains of the RI panel (SHR/Ola and BN-Lx/Cub; 70–90 days old) using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA, USA). Samples were precipitated with sodium acetate to further purify and concentrate DNA. Quantity and quality of DNA samples were determined with a Nanodrop (Thermo Fisher Scientific, Wilmington, DE, USA) and Agilent BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA, USA), respectively. One microgram of genomic DNA in 53 μL of 1× Tris-EDTA was sheared using the S220 Covaris Instrument (Thermo Fisher Scientific). A 300-bp peak was targeted using a duty factor of 10%, peak incident power of 140, 200 cycles per burst and 80 s in duration at 6 °C. One microgram of sheared DNA was then used for sequencing library construction. The Illumina TruSeq DNA Kit (Illumina, San Diego, CA, USA) was used to prepare each library in accordance with the manufacturer's instructions. The DNA in the libraries was quantified using an Invitrogen Qubit Fluorometer (Life Technologies, Grand Island, NY, USA) and an Agilent BioAnalyzer 2100. In total, 5 pmol of each library was sequenced per individual lane using 100 cycle paired-end reads on an Illumina cBot and HiSeq2000 (Illumina) in accordance with the manufacturer's instructions. Each library was sequenced in duplicate in two lanes on a V3 flow cell. Paired-end 100-nucleotide Illumina reads were trimmed to 80 nucleotides. The reads were aligned to the RGSC 5.0/rn5 version of the rat genome using BOWTIE2 [70]. SNP and small indel calls were made using a samtools/bcftools [71] pipeline and were filtered for quality (quality score ≥ 10 and supported by ≥ 3 quality reads) and homozygosity (SNPs/indels with heterozygous calls were discarded).

RNA-Seq

RNA-Seq was performed on two separate RNA fractions: polyA+-selected RNA and ribosomal RNA-depleted total RNA. Total RNA was isolated from brain samples of three rats per progenitor strain (SHR/Ola and BN-Lx/Cub; 70–90 days old) using either the RNeasy Midi Kit with additional clean-up using the RNeasy Mini Kit (Qiagen) for the ribosomal RNA-depleted total RNA preparation or the miRNeasy Mini and RNeasy MinElute Cleanup Kits (Qiagen) for the polyA+ RNA preparation, in accordance with the manufacturer's instructions. The RNeasy Midi Kit protocol isolates and purifies large RNAs (> 200 nucleotides) only. The miRNeasy Mini and RNeasy MinElute Cleanup Kits separate the total RNA into a large RNA fraction (> 200 nucleotides) and a small RNA (< 200 nucleotides) fraction. The small RNA fraction was analyzed separately (data available at http://phenogen.ucdenver.edu), although only the results from the large RNA fraction are reported here. The quality of extracted total RNA (> 200 nucleotides) was assessed on an Agilent Bioanalyzer. Ribosomal RNA was depleted from total RNA (> 200 nucleotides) using the Ribo-Zero Magnetic Kit (Epicentre Biotechnologies, Madison, WI, USA) in accordance with the manufacturer's instructions. The polyA+ RNA was isolated using oligo-dT magnetic beads.

RNA-seq libraries prepared from the polyA+ fraction were constructed using the Illumina TruSeq RNA Sample Preparation kit from 1 μg of RNA in accordance with the manufacturer's instructions. Library quality was assessed using the Agilent Bioanalyzer. For sequencing on the Illumina HiSeq2000, samples were multiplexed over three lanes of the flowcell (two lanes with three samples each and one lane with all six samples).

For the total RNA (ribosomal-RNA depleted RNA) sequencing, libraries were constructed using the Illumina TruSeq RNA Sample Preparation kit at the elution-fragmentation-priming step, in accordance with the manufacturer's instructions. Library quality was assessed using the Agilent Bioanalyzer. Six samples were sequenced using an Illumina HiSeq2000 over five lanes (three lanes with two samples per lane and two lanes with three samples per lane; each sample was included in two lanes).

Prior to alignment, reads were de-multiplexed and read fragments were trimmed for adaptors and for quality (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore). Reads were eliminated if the trimmed length of either read fragment was < 20 nucleotides. Reads were aligned with their respective strain-specific genomes derived from our DNA sequencing using BOWTIE2/TOPHAT suite of tools [72] with the default settings.

Transcriptome reconstruction

A genome-guided transcriptome reconstruction was executed for each progenitor strain using data from the total RNA preparation and the polyA+ fraction separately with the CUFFLINKS algorithm and software [73]. Prior to merging the transcriptomes, 'high confidence' transcripts were identified. A 'high confidence' transcript had an estimated FPKM (fragments·kb transcript^{-1}·million fragments mapped^{-1}) > 1 within at least one strain, and the transcript was longer than 350 nucleotides. High confidence transcripts were merged across strains and across the RNA preparation methods into one transcriptome, using cuffmerge from the CUFFLINKS suite [74]. The merged transcriptome was compared with both the Ensembl database (Rnor_5.0.71) and the RefSeq rat database (RGSC 5.0/rn5) to determine overlap with annotated genes using cuffcompare [74].

Filtering probes and constructing clusters

Individual probe sequences from the Affymetrix Rat Exon 1.0 ST Array were retrieved from the Affymetrix website (http://www.affymetrix.com) and aligned with the RGSC 5.0/rn5 version of the rat genome using the BLAT algorithm [75]. Probes were eliminated if their sequence did not align perfectly to the reference genome or if their sequence aligned perfectly to multiple places in the genome. Probes were also eliminated: (a) if the region of the genome to which they aligned harbored a SNP or small indel between either of the progenitor strains and the BN reference genome (via DNA-Seq) and (b) if <3 probes remained in the probe set after certain probes were eliminated. Probe sets were summarized into probe set clusters based on the transcriptome reconstruction (via our RNA-Seq results). Both 'isoform-level' and 'gene-level' clusters were generated. The rationale for generating isoform-level clusters was to determine whether only a specific isoform of a gene was associated with alcohol consumption. However, because many isoforms were not interrogated by a probe set that was unique to that isoform, we also examined gene-level clusters. When there is a highly expressed isoform, the gene-level cluster will capture its expression levels. For the gene-level analysis, all probe sets that were contained completely within an exon or UTR of a gene expressed in the brain, and did not overlap another gene, were summarized into a gene cluster. For the isoform-level analysis, probe sets were included in an isoform cluster if they aligned with a region of an exon or UTR of a particular isoform that did not overlap any other isoforms or genes.

Detailed methods for identification of candidate genes associated with a predisposition to alcohol preference/consumption

Animals

Alcohol-naïve male rats (60–90 days old) from six separate pairs of lines selectively bred for either high or low alcohol preference were used for our studies. Brain tissues were received from five animals for each line from populations in Indiana, USA (high alcohol-drinking 1 and low alcohol-drinking 1, HAD1/LAD1; high alcohol-drinking 2 and low alcohol-drinking 2, HAD2/LAD2; and alcohol-preferring and alcohol-nonpreferring, P/NP) [76]; five animals from each line from Helsinki, Finland (Alko alcohol and Alko non-alcohol, AA/ANA) [77]; five animals from each line from Cagliari, Italy (Sardinian alcohol-preferring and Sardinian alcohol-nonpreferring, sP/sNP) [78]; and five animals from each line from the University of Chile (UChB/UChA) [79].

Male rats from the HXB/BXH RI panel were also used for these studies. These rats were developed from an intercross between two inbred strains, the Wistar origin spontaneously hypertensive rat (SHR/Ola) and a Brown Norway congenic (BN-Lx/Cub), by Drs Michal Pravenec and Vladimir Kren (Institute of Biology of Charles University and Institute of Physiology of the Czech Academy of Sciences, Prague, Czech Republic). The rats were rederived and maintained by Dr Morton Printz (University of California, San Diego, CA, USA). The RI strains were bred in a gender reciprocal manner, providing strains that differ in the source of mitochondrial DNA and the Y chromosome (HXB and BXH strains) [32].

RNA expression estimates

Total RNA was extracted from individual brains of five male rats per selected line or four male rats per RI strain (21 strains; 70–90 days old) using the RNeasy Midi kit (Qiagen) and the RNeasy Mini kit (Qiagen) for cleanup. cDNA from the brain of each individual rat was hybridized to a separate Affymetrix GeneChip® Rat Exon 1.0 ST array (Affymetrix). Arrays were processed in accordance with the manufacturer's instructions. All processed array data were examined for quality using the tools outlined in detail at http://phenogen.ucdenver.edu.

Gene-level expression estimates and isoform-specific expression estimates were derived using the probe masks described above and the RMA algorithm [80] implemented in the Affymetrix Power Tools (www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx). Expression data were also subjected to a batch effects adjustment using the Combat algorithm [81]. After batch effects adjustment, both individual samples and strain means (RI panel only) were examined for outliers using hierarchical clustering. We chose a criterion for a gene/isoform cluster that at least 5% of samples had to have expression levels above background to include the gene/isoform cluster in further analyses. The threshold of 5% of samples was chosen to ensure that genes/isoforms expressed exclusively in one strain/line were included [68]. Detection above background was determined using the DABG *P* value calculated within the Affymetrix Power Tools suite. The expression value of a gene/isoform cluster for an individual sample was considered to be 'detected above background' if the DABG *P* value was < 0.0001. This threshold is more stringent than the threshold recommended by Affymetrix [82] ($P < 0.05$), although using the recommended criterion would have resulted in a high false positive rate; for example, a probability of 58% that at least 5% of samples (three out of 60) would have a DABG *P* value < 0.05 when none of the expression values are above background. However, the probability of at least 5% of samples (three out of 60) will have a DABG

P value < 0.0001 when none of the samples are expressed above background is $< 1 \times 10^{-7}$.

Selected lines meta-analysis

To determine differential expression of genes/isoforms in the selected lines, a random effects meta-analysis was implemented in SAS (SAS Institute Inc., Cary, NC, USA) using PROC MIXED where the random effect was the selected line pair. For each gene/isoform cluster, two models were evaluated: one that allowed the variance within a selected line pair to vary across pairs and one that constrained the variance to be the same within each pair. The P value from the model with the smaller Akaike information criterion was used to determine differential expression. These P values were adjusted for multiple comparisons across genes and isoforms using a FDR [83]. For a gene/isoform cluster to be associated with drinking in the meta-analysis of the six selected line pairs, not only did it have to be significantly associated with alcohol preference after multiple testing correction (FDR < 0.05), but also the direction of the expression difference in individual selected line pairs had to match (e.g. the higher preferring line had higher expression levels, across all selected line pairs with minimal statistical evidence for a detectable difference; $P < 0.05$ for the individual line pair). The differential expression estimates reported for individual selected line pairs were derived using least squares estimates from the full model.

Weighted gene co-expression network analysis in RI panel

An unsigned weighted gene co-expression network analysis was executed for the HXB/BXH RI panel to identify gene co-expression modules and isoform co-expression modules, separately, using the WGCNA package in R [24]. Two parameters were altered from their default setting to allow for the identification of smaller modules: the minimum module size (was set to 5) and the deepSplit parameter (was set to 4). The Pearson correlation coefficient calculated between gene/isoform clusters was used to generate the network. The model fitting index proposed by Zhang and Horvath [84] was used to determine the appropriate soft thresholding power. A soft-thresholding power of 7 was sufficient for both networks.

Modules associated with alcohol consumption/preference

Data on alcohol consumption were gathered on male rats (70–100 days old at the start of study) from 23 HXB/BXH RI strains and the two progenitor strains at the University of California (San Diego, CA, USA). The number of rats per strain ranged from nine to 12, with 242 total rats being utilized to measure alcohol consumption. In the first week (week 0) of treatment, rats were given 10% ethanol as their only choice of fluid. For the next 7 weeks, the rats were given a choice of two bottles: one with water and one with a 10% (v/v) ethanol solution. For the present study, we used alcohol consumption data from the second week of the two-bottle choice paradigm to match our previous research with this phenotype [19,66]. These studies were performed in accordance with the guidelines in the NIH Guide for the Care and Use of Laboratory Animals, and were approved by the University of California, San Diego Institutional Animal Care and Use Committee.

Initially, co-expression modules were evaluated for association with alcohol consumption using a P value that combined a correlation analysis of the module's eigengene with alcohol consumption (week 2) from the HXB/BXH RI panel with an analysis that evaluated the module based on enrichment of genes that were identified as differentially expressed in the selected lines meta-analysis (outlined earlier). The P values from these two analyses were combined using Fisher's method and modules were retained if their combined P value was < 0.01.

The list of candidate co-expression modules was further reduced by only considering modules with a significant eigengene QTL that overlaps a behavioral QTL for alcohol consumption in the HXB/BXH RI panel. The marker set used for QTL analysis in the HXB/BXH rats was derived from the SNPs genotyped by the STAR consortium (http://www.snp-star.eu) [85]. The locations of SNPs were converted to the RGSC 5.0/rn5 version of the rat genome and their genotypes were examined in detail for quality as outlined in Vanderlinden *et al.* [66]. QTLs for alcohol consumption and for module eigengenes in the HXB/BXH panel were calculated using a marker regression on strain means (21 RI strains with both genotype and alcohol consumption data). Results are reported for individual marker/phenotype (or eigengene) associations using a LOD score (i.e. the log base 10 of the likelihood ratio that compares a model that includes a genotype effect for that marker versus a model without a genotype effect). Empirical genome-wide P values were calculated for all QTL analyses using 1000 permutations [86]. QTLs with empirical genome-wide P values < 0.05 were considered statistically significant and QTLs with empirical genome-wide P values < 0.63 were considered suggestive based on guidelines presented by Lander and Kruglyak [27] and adopted by many (e.g. the Complex Trait Consortium) [87]. Bayesian credible intervals were calculated for alcohol consumption QTLs using methods outlined previously [88]. Confidence intervals for eigengene QTLs were calculated using the bootstrap method described in Visscher *et al.* [89]. Alcohol consumption QTL analyses and graphics were generated using the R/qtl package in R [90]. Because of the number of eigengenes analyzed, eigengene QTLs were calculated using QTLREAPER (http://qtlreaper.sourceforge.net).

Candidate genes

To identify not only individual genes/isoforms related to alcohol consumption, but also functional pathways, we gathered a list of annotated candidate genes from both the co-expression module associated with alcohol consumption and the differentially expressed genes/isoforms from the selected lines meta-analysis. Genes/isoforms from the candidate co-expression module were filtered for independent correlation with alcohol consumption in the HXB/BXH panel ($P < 0.05$) and were combined with the ten genes/isoforms with the most significant association with alcohol consumption in the selected-lines meta-analysis. The purpose of putting together a list of candidate genes/isoforms was to be able to systematically identify shared functional pathways among genes with the most evidence of association with a predisposition to alcohol preference/consumption. This list is meant to be representative rather than exhaustive.

Summary

The results of the present study show that different selectively bred rat lines and RI strains may display different combinations of differentially expressed genes influencing the risk for alcohol drinking. However, there are common functional pathways that are involved in all models that we have studied. Because high levels of alcohol consumption represent a risk factor for alcohol addiction [5], the neurobiological systems identified in our studies (e.g. neuroinflammation, energy metabolism, cell—cell communication) can serve to focus future studies with humans on the genetic predisposition for high alcohol consumption and by extrapolation [5] for alcohol dependence.

Acknowledgements

This work was supported in part by NIAAA/NIH [R24AA013162 (BT); U01AA016649 INIA project (PLH); U01AA016663 INIA project (BT); AA006420 (GK, HNR); AAU01 Developmental Grant-INIA Project (MPP); T32AA007464 (SLF)]; NHLBI/NIH [HL35018 (MPP)]; Banbury Fund (BT); and the Pearson Center for Alcoholism and Addiction Research (GK). The authors are grateful for the expert technical assistance with microarrays and RNA-Seq provided by Yinni Yu and Adam Chapman and for the expert technical assistance with the behavioral studies provided by Laura Breen and Joseph Gatewood.

Author contributions

LMS assisted with the experimental design, developed and performed data analysis, helped interpret analyzed data, and assisted with the preparation of manuscript. SCF performed genome sequencing and reconstructed rat strain specific genomes. LAV performed statistical analysis for QTL locations. YI and LT recovered brains from UChB and UChA rats and provided tissue for analysis. GC recovered brains from sP and sNP rats and provided tissue for analysis. KK recovered brains from AA and ANA rats and provided tissue for analysis. RLB recovered brains from HAD_1, HAD_2, P and LAD_1, LAD_2, NP rats and provided tissue for analysis. MPP bred the rats of the HXB/BXH RI panel and provided rats for behavioral, genomic and transcriptome studies. PF assisted with the breeding of the HXB/BXH rats, recovered brain tissue from all strains, and assisted with the behavioral studies. GK generated the experimental design for measuring alcohol consumption in the HXB/BXH rats and analyzed data. HNR performed the alcohol consumption studies with HXB/BXH rats, and collected and analyzed data. JL managed the supercomputer systems for data analysis and generated solutions for streamlining analysis. PLH assisted with the data analysis, as well as the interpretation and writing of all aspects of the manuscript. BT conceived the experimental goals and design and data analysis, performed data analysis and interpretation, and wrote the manuscript. All authors read and approved the final manuscript submitted for publication.

References

1 Tabakoff B, Saba L, Kechris K, Hu W, Bhave SV, Finn DA, Grahame NJ & Hoffman PL (2008) The genomic determinants of alcohol preference in mice. *Mamm Genome* **19**, 352–365.

2 Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA *et al.* (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37**, 233–242.

3 Kang HP, Yang X, Chen R, Zhang B, Corona E, Schadt EE & Butte AJ (2012) Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2 diabetes. *Diabetologia* **55**, 2205–2213.

4 Walter NA, McWeeney SK, Peters ST, Belknap JK, Hitzemann R & Buck KJ (2007) SNPs matter: impact on detection of differential expression. *Nat Methods* **4**, 679–680.

5 Dawson DA & Grant BF (2011) The 'gray area' of consumption between moderate and risk drinking. *J Stud Alcohol Drugs* **72**, 453–458.

6 Grant JD, Agrawal A, Bucholz KK, Madden PA, Pergadia ML, Nelson EC, Lynskey MT, Todd RD, Todorov AA, Hansell NK *et al.* (2009) Alcohol consumption indices of genetic risk for alcohol dependence. *Biol Psychiatry* **66**, 795–800.

7 Kendler KS, Myers J, Dick D & Prescott CA (2010) The relationship between genetic influences on alcohol dependence and on patterns of alcohol consumption. *Alcohol Clin Exp Res* **34**, 1058–1065.

8 Kapoor M, Wang JC, Wetherill L, Le N, Bertelsen S, Hinrichs AL, Budde J, Agrawal A, Bucholz K, Dick D *et al.* (2013) A meta-analysis of two genome-wide association studies to identify novel loci for maximum number of alcoholic drinks. *Hum Genet* **132**, 1141–1151.

9 Murphy JM, Stewart RB, Bell RL, Badia-Elder NE, Carr LG, McBride WJ, Lumeng L & Li TK (2002) Phenotypic and genotypic characterization of the Indiana University rat lines selectively bred for high and low alcohol preference. *Behav Genet* **32**, 363–388.

10 Heath AC, Meyer J, Jardine R & Martin NG (1991) The inheritance of alcohol consumption patterns in a general population twin sample: II. Determinants of consumption frequency and quantity consumed. *J Stud Alcohol* **52**, 425–433.

11 Swan GE, Carmelli D, Rosenman RH, Fabsitz RR & Christian JC (1990) Smoking and alcohol consumption in adult male twins: genetic heritability and shared environmental influences. *J Subst Abuse* **2**, 39–50.

12 Young-Wolff KC, Enoch MA & Prescott CA (2011) The influence of gene-environment interactions on alcohol consumption and alcohol use disorders: a comprehensive review. *Clin Psychol Rev* **31**, 800–816.

13 Belknap JK & Atkins AL (2001) The replicability of QTLs for murine alcohol preference drinking behavior across eight independent studies. *Mamm Genome* **12**, 893–899.

14 Ehlers CL, Walter NA, Dick DM, Buck KJ & Crabbe JC (2010) A comparison of selected quantitative trait loci associated with alcohol use phenotypes in humans and mouse models. *Addict Biol* **15**, 185–199.

15 Falconer DS & Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longman Group Ltd., Essex,UK.

16 McBride WJ, Kimpel MW, McClintick JN, Ding ZM, Hyytia P, Colombo G, Edenberg HJ, Lumeng L & Bell RL (2012) Gene expression in the ventral tegmental area of 5 pairs of rat lines selectively bred for high or low ethanol consumption. *Pharmacol Biochem Behav* **102**, 275–285.

17 McBride WJ, Kimpel MW, McClintick JN, Ding ZM, Hyytia P, Colombo G, Liang T, Edenberg HJ, Lumeng L & Bell RL (2013) Gene expression within the extended amygdala of 5 pairs of rat lines selectively bred for high or low ethanol consumption. *Alcohol* **47**, 517–529.

18 Saba LM, Bennett B, Hoffman PL, Barcomb K, Ishii T, Kechris K & Tabakoff B (2011) A systems genetic analysis of alcohol drinking by mice, rats and men: influence of brain GABAergic transmission. *Neuropharmacology* **60**, 1269–1280.

19 Tabakoff B, Saba L, Printz M, Flodman P, Hodgkinson C, Goldman D, Koob G, Richardson HN, Kechris K, Bell RL *et al.* (2009) Genetical genomic determinants of alcohol consumption in rats and humans. *BMC Biol* **7**, 70.

20 Bell RL, Sable HJ, Colombo G, Hyytia P, Rodd ZA & Lumeng L (2012) Animal models for medications development targeting alcohol abuse using selectively bred rat lines: neurobiological and pharmacological validity. *Pharmacol Biochem Behav* **103**, 119–155.

21 Gora-Maslak G, McClearn GE, Crabbe JC, Phillips TJ, Belknap JK & Plomin R (1991) Use of recombinant inbred strains to identify quantitative trait loci in psychopharmacology. *Psychopharmacology* **104**, 413–424.

22 Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S & Geschwind DH (2008) Functional organization of the transcriptome in human brain. *Nat Neurosci* **11**, 1271–1282.

23 Oldham MC, Horvath S & Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci USA* **103**, 17973–17978.

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

25 Vanderlinden LA, Saba LM, Kechris K, Miles MF, Hoffman PL & Tabakoff B (2013) Whole brain and brain regional coexpression network interactions associated with predisposition to alcohol consumption. *PLoS One* **8**, e68878.

26 Kren V (1975) Genetics of the polydactyly-luxate syndrome in the Norway rat, Rattus norvegicus. *Acta Univ Carol Med Monogr* **68**, 1–103.

27 Lander E & Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11**, 241–247.

28 Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR & Horvath S (2011) Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* **12**, 322.

29 Alam M, Coulet A, Napoli A & Smail-Tabbone M. Formal concept analysis applied to transcriptomic data. Proceedings of Conference: What can FCA do for Artificial Inteligence (FCA4A). ECAI 2012: http://hal.inria.fr/hal-00760993.

30 Yu Y, Fuscoe JC, Zhao C, Guo C, Jia M, Qing T, Bannon DI, Lancashire L, Bao W, Du T *et al.* (2014) A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat Commun* **5**, 3230.

31 Hermsen R, de Ligt J, Spee W, Blokzijl F, Schafer S, Adami E, Boymans S, Flink S, van Boxtel R, van der Weide RH *et al.* Genomic landscape of rat strain and substrain variation. *BMC Genom* **16**, 357.

32 Printz MP, Jirout M, Jaworski R, Alemayehu A & Kren V (2003) Genetic Models in Applied Physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. *J Appl Physiol* **94**, 2510–2522.

33 Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428.

34 Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, Ruff TG, Milligan SB, Lamb JR, Cavet G *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.

35 Sheridan GK & Murphy KJ (2013) Neuron-glia crosstalk in health and disease: fractalkine and CX3CR1 take centre stage. *Open Biol* **3**, 130181.

36 Kaadige MR, Looper RE, Kamalanaadhan S & Ayer DE (2009) Glutamine-dependent anapleurosis dictates glucose uptake and cell growth by regulating MondoA transcriptional activity. *Proc Natl Acad Sci USA* **106**, 14878–14883.

37 Thompson JL & Shuttleworth TJ (2011) Orai channel-dependent activation of phospholipase C-delta: a novel mechanism for the effects of calcium entry on calcium oscillations. *J Physiol* **589**, 5057–5069.

38 Coddou C, Yan Z, Obsil T, Huidobro-Toro JP & Stojilkovic SS (2011) Activation and regulation of purinergic P2X receptor channels. *Pharmacol Rev* **63**, 641–683.

39 Malorni W, Farrace MG, Rodolfo C & Piacentini M (2008) Type 2 transglutaminase in neurodegenerative diseases: the mitochondrial connection. *Curr Pharm Des* **14**, 278–288.

40 Dong B & Silverman RH (1995) 2-5A-dependent RNase molecules dimerize during activation by 2-5A. *J Biol Chem* **270**, 4133–4137.

41 Jha BK, Polyakova I, Kessler P, Dong B, Dickerman B, Sen GC & Silverman RH (2011) Inhibition of RNase L and RNA-dependent protein kinase (PKR) by sunitinib impairs antiviral innate immunity. *J Biol Chem* **286**, 26319–26326.

42 Rossol M, Pierer M, Raulien N, Quandt D, Meusch U, Rothe K, Schubert K, Schöneberg T, Schaefer M, Krügel U *et al.* (2012) Extracellular Ca2+ is a danger signal activating the NLRP3 inflammasome through G protein-coupled calcium sensing receptors. *Nat Commun* **3**, 1329.

43 Tsuda M, Toyomitsu E, Kometani M, Tozaki-Saitoh H & Inoue K (2009) Mechanisms underlying fibronectin-induced up-regulation of P2X4R expression in microglia: distinct roles of PI3K-Akt and MEK-ERK signalling pathways. *J Cell Mol Med* **13**, 3251–3259.

44 Nurminskaya MV & Belkin AM (2012) Cellular functions of tissue transglutaminase. *Int Rev Cell Mol Biol* **294**, 1–97.

45 Missler M & Sudhof TC (1998) Neurexophilins form a conserved family of neuropeptide-like glycoproteins. *J Neurosci* **18**, 3630–3638.

46 Pettem KL, Yokomaku D, Luo L, Linhoff MW, Prasad T, Connor SA, Siddiqui TJ, Kawabe H, Chen F, Zhang L *et al.* (2013) The specific alpha-neurexin interactor calsyntenin-3 promotes excitatory and inhibitory synapse development. *Neuron* **80**, 113–128.

47 Foster MW, Thompson JW, Forrester MT, Sha Y, McMahon TJ, Bowles DE, Moseley MA & Marshall HE (2013) Proteomic analysis of the NOS2 interactome in human airway epithelial cells. *Nitric Oxide* **34**, 37–46.

48 Tada H, Okano HJ, Takagi H, Shibata S, Yao I, Matsumoto M, Saiga T, Nakayama KI, Kashima H, Takahashi T *et al.* (2010) Fbxo45, a novel ubiquitin ligase, regulates synaptic activity. *J Biol Chem* **285**, 3840–3849.

49 Bhogaraju S, Cajanek L, Fort C, Blisnick T, Weber K, Taschner M, Mizuno N, Lamla S, Bastin P, Nigg EA *et al.* (2013) Molecular basis of tubulin transport within the cilium by IFT74 and IFT81. *Science* **341**, 1009–1012.

50 Besschetnova TY, Kolpakova-Hart E, Guan Y, Zhou J, Olsen BR & Shah JV (2010) Identification of signaling pathways regulating primary cilium length and flow-mediated adaptation. *Curr Biol* **20**, 182–187.

51 Berbari NF, Pasek RC, Malarkey EB, Yazdi SM, McNair AD, Lewis WR, Nagy TR, Kesterson RA & Yoder BK (2013) Leptin resistance is a secondary consequence of the obesity in ciliopathy mutant mice. *Proc Natl Acad Sci USA* **110**, 7796–7801.

52 Lippai D, Bala S, Petrasek J, Csak T, Levin I, Kurt-Jones EA & Szabo G (2013) Alcohol-induced IL-1beta in the brain is mediated by NLRP3/ASC inflammasome activation that amplifies neuroinflammation. *J Leukoc Biol* **94**, 171–182.

53 Suokas A, Forsander O & Lindros K (1984) Distribution and utilization of alcohol-derived acetate in the rat. *J Stud Alcohol* **45**, 381–385.

54 Carmichael FJ, Israel Y, Crawford M, Minhas K, Saldivia V, Sandrin S, Campisi P & Orrego H (1991) Central nervous system effects of acetate: contribution to the central effects of ethanol. *J Pharmacol Exp Ther* **259**, 403–408.

55 Muir D, Berl S & Clarke DD (1986) Acetate and fluoroacetate as possible markers for glial metabolism in vivo. *Brain Res* **380**, 336–340.

56 Schousboe A, Sickmann HM, Bak LK, Schousboe I, Jajo FS, Faek SA & Waagepetersen HS (2011) Neuron-glia interactions in glutamatergic neurotransmission: roles of oxidative and glycolytic adenosine triphosphate as energy source. *J Neurosci Res* **89**, 1926–1934.

57 Volkow ND, Kim SW, Wang GJ, Alexoff D, Logan J, Muench L, Shea C, Telang F, Fowler JS, Wong C *et al.* (2013) Acute alcohol intoxication decreases glucose metabolism but increases acetate uptake in the human brain. *NeuroImage* **64**, 277–283.

58 Pawlosky RJ, Kashiwaya Y, Srivastava S, King MT, Crutchfield C, Volkow N, Kunos G, Li TK & Veech RL (2010) Alterations in brain glucose utilization accompanying elevations in blood ethanol and acetate concentrations in the rat. *Alcohol Clin Exp Res* **34**, 375–381.

59 Mosher KI & Wyss-Coray T (2014) Microglial dysfunction in brain aging and Alzheimer's disease. *Biochem Pharmacol* **88**, 594–604.

60 Rosenblat JD, Cha DS, Mansur RB & McIntyre RS (2014) Inflamed moods: a review of the interactions between inflammation and mood disorders. *Prog Neuropsychopharmacol Biol Psychiatry* **53**, 23–34.

61 Crews FT, Zou J & Qin L (2011) Induction of innate immune genes in brain create the neurobiology of addiction. *Brain Behav Immun* **25** (Suppl 1), S4–S12.

62 Mayfield J, Ferguson L & Harris RA (2013) Neuroimmune signaling: a key component of alcohol abuse. *Curr Opin Neurobiol* **23**, 513–520.

63 Qin L, He J, Hanes RN, Pluzarev O, Hong JS & Crews FT (2008) Increased systemic and brain cytokine production and neuroinflammation by endotoxin following ethanol treatment. *J Neuroinflammation* **5**, 10.

64 Blednov YA, Benavidez JM, Geil C, Perra S, Morikawa H & Harris RA (2011) Activation of inflammatory signaling by lipopolysaccharide produces a prolonged increase of voluntary alcohol intake in mice. *Brain Behav Immun* **25** (Suppl 1), S92–S105.

65 Soliman ML, Combs CK & Rosenberger TA (2013) Modulation of inflammatory cytokines and mitogen-activated protein kinases by acetate in primary astrocytes. *J Neuroimmune Pharmacol* **8**, 287–300.

66 Vanderlinden LA, Saba LM, Printz MP, Flodman P, Koob G, Richardson HN, Hoffman PL & Tabakoff B (2014) Is the alcohol deprivation effect genetically mediated? Studies with HXB/BXH recombinant inbred rat strains. *Alcohol Clin Exp Res* **38**, 2148–2157.

67 Jiang L, Gulanski BI, De Feyter HM, Weinzimer SA, Pittman B, Guidone E, Koretski J, Harman S, Petrakis IL, Krystal JH *et al.* (2013) Increased brain uptake and oxidation of acetate in heavy drinkers. *J Clin Investig* **123**, 1605–1614.

68 McClintick JN, Xuei X, Tischfield JA, Goate A, Foroud T, Wetherill L, Ehringer MA & Edenberg HJ (2013) Stress-response pathways are altered in the hippocampus of chronic alcoholics. *Alcohol* **47**, 505–515.

69 Atanur SS, Diaz AG, Maratou K, Sarkis A, Rotival M, Game L, Tschannen MR, Kaisaki PJ, Otto GW, Ma MC *et al.* (2013) Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell* **154**, 691–703.

70 Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359.

71 Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993.

72 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R & Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36.

73 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ & Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515.

74 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL & Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578.

75 Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 656–664.

76 Li TK, Lumeng L & Doolittle DP (1993) Selective breeding for alcohol preference and associated responses. *Behav Genet* **23**, 163–170.

77 Eriksson K (1968) Genetic selection for voluntary alcohol consumption in the albino rat. *Science* **159**, 739–741.

78 Colombo G. ESBRA-Nordmann 1996 Award Lecture: ethanol drinking behaviour in Sardinian alcohol-preferring rats. Alcohol and Alcoholism; Oxford, Oxfordshire 1997, July.

79 Quintanilla ME, Israel Y, Sapag A & Tampier L (2006) The UChA and UChB rat lines: metabolic and genetic differences influencing ethanol intake. *Addict Biol* **11**, 310–323.

80 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B & Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15.

81 Johnson WE, Li C & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127.

82 Affymetrix (2006) Identifying and validating alternative splicing events. http://mediaaffymetrixcom/support/technical/technotes/id_altsplicingevents_technotepdf.

83 Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* **57**, 289–300.

84 Zhang B & Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17.

85 STAR Consortium, Saar K, Beck A, Bihoreau M-T, Birney E & Brocklebank D (2008) SNP and haplotype mapping for genetic analysis in the rat. *Nat Genet* **40**, 560–566.

86 Churchill GA & Doerge RW (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

87 Abiola O, Angel JM, Avner P, Bachmanov AA, Belknap JK, Bennett B, Blankenhorn EP, Blizard DA, Bolivar V, Brockmann GA *et al.* (2003) The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet* **4**, 911–916.

88 Sen S & Churchill GA (2001) A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387.

89 Visscher PM, Thompson R & Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.

90 Broman KW, Wu H, Sen S & Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Fig. S1. Complete workflow from generating the probe mask and defining transcript clusters to the identification of functional categories represented in candidate genes for predisposition to alcohol consumption/preference.

Fig. S2. Distribution of average daily alcohol consumption using the two-bottle 24-h access paradigm in the HXB/BXH recombinant inbred rat panel.

Fig. S3. Distribution of module size in WGCNA.

Fig. S4. Comparison of LOD profiles between alcohol consumption and module eigengenes.

Doc. S1. Candidate module hub gene.

Doc. S2. Individual gene reports.

Doc. S3. Functional analysis.