nature
genetics

Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function

Elissa J Chesler[1,5], Lu Lu[1,5], Siming Shou[1], Yanhua Qu[1], Jing Gu[1], Jintao Wang[1], Hui Chen Hsu[2], John D Mountz[2], Nicole E Baldwin[3], Michael A Langston[3], David W Threadgill[4], Kenneth F Manly[1] & Robert W Williams[1]

Patterns of gene expression in the central nervous system are highly variable and heritable. This genetic variation among normal individuals leads to considerable structural, functional and behavioral differences. We devised a general approach to dissect genetic networks systematically across biological scale, from base pairs to behavior, using a reference population of recombinant inbred strains. We profiled gene expression using Affymetrix oligonucleotide arrays in the BXD recombinant inbred strains, for which we have extensive SNP and haplotype data. We integrated a complementary database comprising 25 years of legacy phenotypic data on these strains. Covariance among gene expression and pharmacological and behavioral traits is often highly significant, corroborates known functional relations and is often generated by common quantitative trait loci. We found that a small number of major-effect quantitative trait loci jointly modulated large sets of transcripts and classical neural phenotypes in patterns specific to each tissue. We developed new analytic and graph theoretical approaches to study shared genetic modulation of networks of traits using gene sets involved in neural synapse function as an example. We built these tools into an open web resource called WebQTL that can be used to test a broad array of hypotheses.

Differences in mRNA expression are generated by complex, dynamic interactions of environmental factors, cell-cell interactions and heritable genetic variation. The genetic component of variation is due to differences that are produced by *cis*-acting polymorphisms often located in a gene's promoter region[1] and by *trans*-acting variants distributed throughout the genome[2,3]. *Trans*-acting modulators of steady-state mRNA abundance include classical transcription factors, RNA helicases, ribozymes and other proteins involved in transcription, RNA processing and degradation. *Trans*-acting factors also include many non-nuclear proteins that influence gene expression through complex molecular cascades, feedback loops and large-scale networks. For example, polymorphisms in neuronal calcium channels have diverse and often indirect repercussions on numerous downstream neural transcription targets[4]. These polymorphisms exert widespread pleiotropic effects on phenotypes ranging from simple steady-state transcript abundance to complex behaviors.

Detecting genetic covariance across biological scale is a challenge. One solution uses a genetic reference population (GRP) of recombinant inbred (RI) strains, from which diverse phenotypes and genotypes can be collected and reproduced over time by many investigators[5,6]. The BXD RI mapping panel was first generated at the Jackson Laboratory in the mid 1970s, was recently extended to 80 strains[7] and is useful for integrative genomics research. These strains have been used by hundreds of investigators for more than two decades to study the genetics of a wide variety of phenotypes. Quantitative trait loci (QTLs) underlying several of these phenotypes were later cloned, including the saccharin preference locus, *Taste*[8]; the kappa-opioid analgesia locus, *Mc1r*[9], whose homolog is involved in human clinical pain; and the alcohol withdrawal seizure locus, *Mpdz*[10]. A particularly compelling advantage of this RI set is that the two parental strains, C57BL/6J and DBA/2J[11], are sequenced. This greatly increases the efficiency of positional candidate gene evaluation. Finally, the BXD panel has been sufficiently studied so that pleiotropic relations and genetic networks can be efficiently constructed. We built both a phenotype database and companion analytic tools for public use so that phenotypes collected using RI strains can be readily integrated into a growing multi-scale base of knowledge of the mouse.

In the same way that one can identify loci that control differences in brain structure or behavior[12], it is now possible to map upstream modulators for thousands of transcripts systematically using

[1]University of Tennessee Health Science Center, 855 Monroe Avenue, Memphis, Tennessee 38163, USA. [2]Department of Medicine, University of Alabama at Birmingham, 701 S. 19th St., Birmingham, Alabama 35294, USA. [3]Department of Computer Science, University of Tennessee-Knoxville, Knoxville, Tennessee 37996-3450, USA. [4]Department of Genetics, CB# 7264, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to R.W.W. (rwilliam@nb.utmem.edu).

microarrays. *Cis-* and *trans-*acting modulators of transcription can be identified by treating mRNA levels as conventional quantitative traits. In the first study of this type, Brem and colleagues[3] used short oligonucleotide arrays and a genetic cross between two yeast strains to define *cis-* and *trans-*acting QTLs and to map the global transcriptional response to starvation. Schadt[13] and colleagues applied this method to the mouse liver using a C5BL/6J × DBA2J F$_2$ cross, again identifying key transcription regulatory regions and association with a single phenotype. Here, we extended this strategy to cumulative systems biological research by analyzing the extensively phenotyped BXD RI strains. We detected and characterized QTLs that modulate transcription of individual genes and large gene networks in what may be the most complex of mouse tissues, the brain. In an accompanying paper, Bystrykh and colleagues[14] extend the approach to an isolated cell population of hematopoietic stem cells (HSCs).

RESULTS

Genetic variation in gene expression

Genetic differences in transcript abundance across the set of 35 strains are substantial (**Supplementary Table 1** online). Differences from two- to fourfold are common among neurologically relevant transcripts. For example, guanine nucleotide binding protein 1 (*Gnb1*, probe set 94853_at) expression has a fivefold range and a heritability of 75%. The abundance of many transcripts is highly heritable and, therefore, amenable to complex-trait analysis even when expression in the parental strains does not differ significantly (a phenomenon known as transgression). With three replicate arrays per strain, we estimate that the amount of variance accounted for by strain (heritability) has a median of 11% and is as high as 78% across all transcripts. An advantage of using RI strains is that effective heritability is increased by additional replication within strains, rendering practical the genetic dissection of phenotypes with modest heritability[15].

Mapping modulators of gene expression

We mapped large numbers of QTLs that modulate transcript abundance at a conventional genome-wise permutation significance threshold of $P < 0.05$ for each transcript. Whole-genome maps for all traits (**Fig. 1**) can be recomputed using a variety of analytic methods in WebQTL[16], including simple and composite interval mapping, pairwise QTL scans, trait clustering and principal component regression. Peak likelihood ratio statistic (LRS) scores across the genome for each transcript range from ~9 to 83 (corresponding to lod scores of ~2.0–18.0). It is possible to localize QTLs for highly penetrant monogenic (mendelian) phenotypes with lod peaks >6 (LRS = 27.6) to intervals as small as 2–4 Mb (**Fig. 2**). This interval size is amenable to sequence comparison and candidate gene analysis[17].
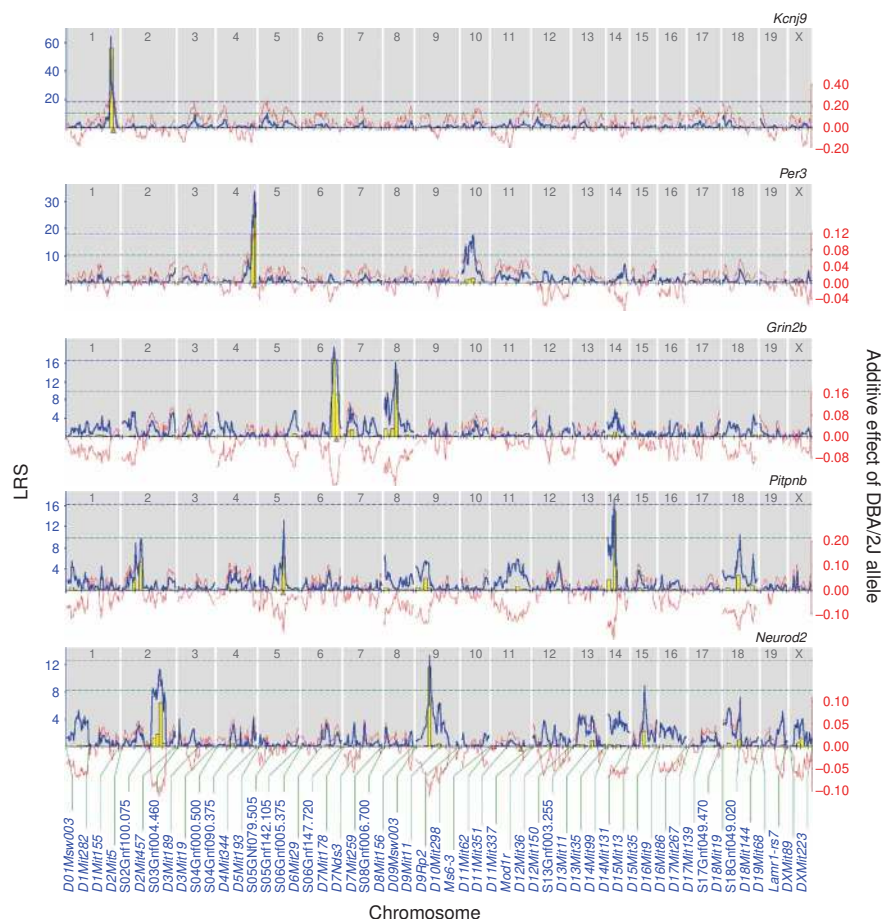
Figure 1 Genome-wide interval mapping for several transcripts from WebQTL, including the *cis*-regulatory QTL for *Kcnj9* (98322_at), *Per3* (102242_at) and *Grin2b* (101312_at) and *trans*-regulatory QTLs for *Pitpnb* (102696_s_at) and *Neurod2* (98808_at). The solid blue line indicates LRS across the genome. A positive additive regression coefficient (red line) indicates that DBA/2J alleles increase trait values, whereas a negative coefficient indicates that C57BL/6J alleles increase trait values. Dashed horizontal lines mark the transcript-specific significance thresholds for genome-wide $P < 0.05$ (significant, blue) and genome-wide $P < 0.63$ (suggestive, green) based on results of 2,000 permutations of the original trait data. The yellow bars indicate the relative frequency of peak LRS at a given location among 2,000 bootstrap resamples.

Resolution will vary across the genome depending on the length of unrecombined haplotype blocks[18].

We applied a permutation test to control the error rate over the whole genome for each single transcript. To control the error rate over the entire set of transcripts, we applied the false-discovery rate (FDR)[19] to these empirical *P* values[20]. A set of 88 QTLs (**Supplementary Table 2** online) met two stringent criteria for statistical significance across the study: low *P* value at the peak LRS for each transcript and high trait heritability. The former is a measure of the strength of association of expression levels to markers; the latter indicates the signal-to-noise ratios in expression estimates. For moderately and highly heritable transcripts (those 608 with >33% variance accounted for by strain), the point estimate of the FDR (q value)[21] is 25% for a genome-wise *P* value of 0.05. This defines 101 significant transcripts, whereas an FDR of 10% (*P* < 0.02) defines 88 significant transcripts. Naturally, the FDR at a given *P* value declines among those transcripts with higher heritabilities, but conservative filtering approaches result in many false negatives.
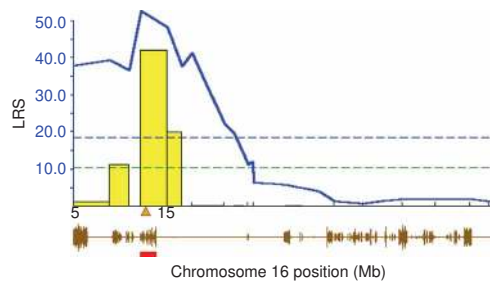
Figure 2 Physical map of chromosome 16 showing *cis*-regulatory locus for expression of a pyridoxil-dependent decarboxylase (160163_at). The location of the polymorphism responsible for variation in mRNA abundance for this transcript was obtained using WebQTL's high-density genetic map. Positional precision is refined by the incorporation of bootstrap analysis (yellow bar) and SNP density analysis (brown). The red bar indicates that the most probable region of the trait relevant polymorphism is at the location of the transcript's coding sequence.
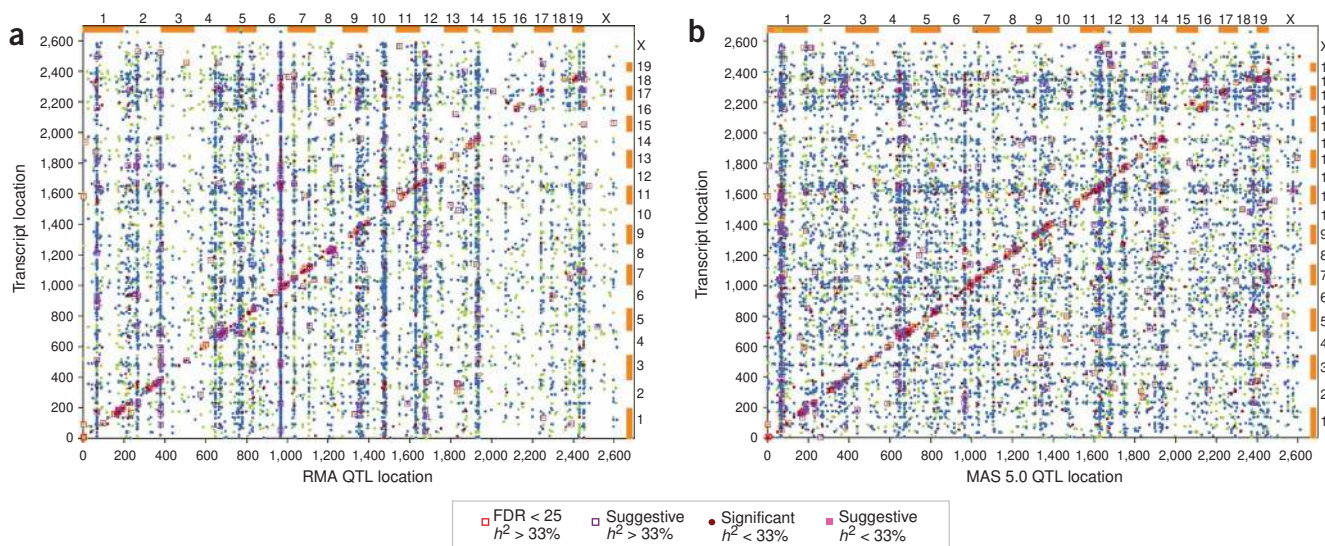
The transcriptome map

Several global properties of transcript modulation in brain are uncovered by plotting the positions of the peak QTLs against the positions of transcripts themselves (**Fig. 3**). First is the presence of a diagonal band of QTLs, located almost precisely at the locations of the transcripts themselves. These *cis*-acting QTLs account for 83 of the 88 QTLs defined at an FDR of 10% (**Supplementary Table 2** online). Second are the vertical bands generated by the coregulation of large numbers of transcripts by single loci. These comodulated transcripts are robust to normalization method (including robust multichip average (RMA), PDNN and MAS 5.0) and are also prominent when transcripts with comparatively low abundance and relatively modest *P* values and heritabilities are plotted. Analysis of HSCs in these same BXD strains[14] identified almost completely different sets of these master *trans*-acting QTLs, indicative of tissue specificity of regulation

of gene expression. Analysis of the mouse liver[13] also identified a different set of *trans*-regulatory QTLs.

### Key modulatory loci control hundreds of transcripts

The seven key *trans*-regulatory QTL bands are located near the following markers on chromosomes 1, 2, 6, 10, 11, 14 and 19: *Mtap2*, *D2Mit200*, *D6Mit150*, *D10Mit42–D10Mit186*, *D11Mit99*, S14Gnf051.890 and *D19Mit13* (**Fig. 4** and **Supplementary Table 3** online). A particularly important regulatory locus is located on chromosome 6 near the marker *D6Mit150* (117.785 Mb). This locus modulates the abundance of ∼1,650 transcripts, more than 10% of all transcripts on the array. Examples of neurologically relevant downstream targets of this locus include *Slc6a1* (161059_at), *Gad1* (103061_at), *Reln* (96591_at), *Adra2b* (99802_at), *Htr4* (95323_at), *Mapk1* (93254_at), *Map3k4* (161007_at), *Mapk6* (103416_at), *Chrng* (95639_at) and *Calm4* (93744_at). The existence of a master modulatory locus immediately raised several questions. Do these downstream targets participate in common cellular functions? Are any of these transcripts *cis*-regulated? Do any of these transcripts contain missense polymorphisms? Do these many transcripts point to a candidate gene? Transcription factors are the main category of transcripts regulated by the *D6Mit150* locus, based on Gene Ontology[22] category representation analysis using the Gene Ontology Tree Machine[23]. These include *Rpo1-4* (161379_at, 162006_r_at and 93620_at), *Hoxb6* (103445_at*)*, *Msx3* (92912_at), *Pax3* (100697_at), *Tcfe2a* (98040_at), *Tead3* (100971_at), *Barx1* (162321_at), *Bach1* (93142_at), *Cdx4* (98347_at), *Dlx4* (98873_at), *Gata6* (104698_at) and *Hes6* (97335_at). The presence of a single locus that simultaneously affects numerous transcription factors is consistent with widespread downstream effects of transcription factor activation.

Several key regulators of transcription are located in the intervals flanking *D6Mit150* (**Fig. 4**). Functional polymorphisms in DNA-binding proteins in the region may regulate the expression of these genes. Examples include *Fbxl14*, *Foxj2* and several zinc-finger proteins


a

b

| FDR < 25 h^2 > 33% | Suggestive h^2 > 33% | Significant h^2 < 33% | Suggestive h^2 < 33% |

Figure 3 Transcriptome maps for 12,422 transcripts. The maps were obtained using RMA (**a**) or MAS 5.0 (**b**) normalization. Chromosome locations are indicated with orange blocks. For traits with more than 33% variance (h^2) accounted for by genetic factors, 101 QTLs are significant at an FDR of 25% (red boxed points in **a**) and 176 have suggestive loci. For the remaining transcription phenotypes (h^2 < 33%), 354 traits have at least one QTL at *P* < 0.05, and 6,351 have at least one suggestive QTL. The diagonal consists of transcripts regulated by polymorphisms that are tightly linked with their own coding sequence. Vertical bands represent *trans*-regulatory QTLs. The less distinct horizontal banding is caused by unequal representation of genes on the Affymetrix array and unequal distribution across the mouse genome. The robustness of *trans*-regulatory bands among traits with low heritability and across normalizations that dampen LRS illustrates that conservative approaches will often produce false negatives.
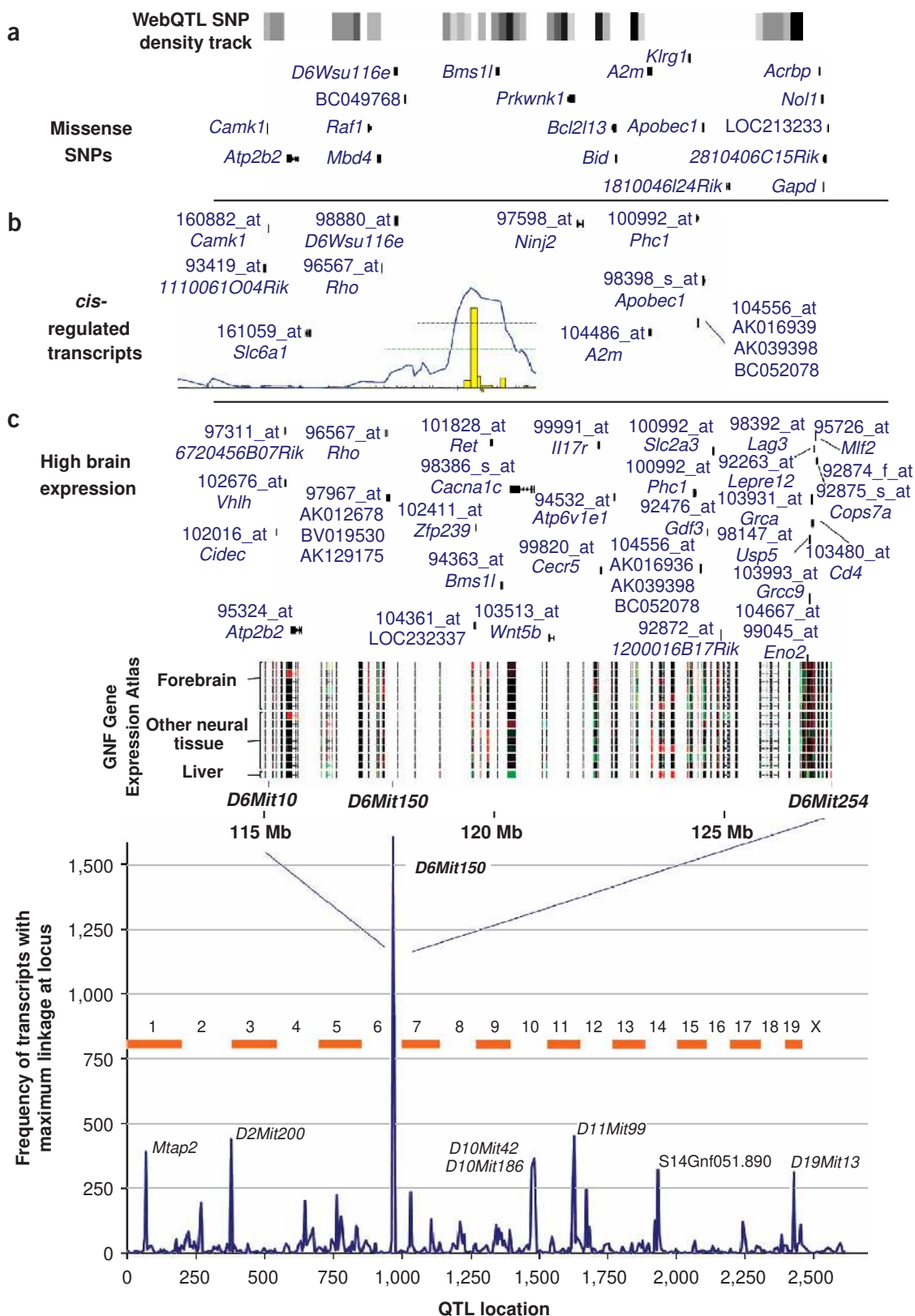
Figure 4 Frequency of transcript abundances with LRS peaks mapping to 5-Mb QTL location bins identify approximately seven key *trans*-regulatory QTLs. Chromosomes are indicated by orange and white bars. The *trans*-regulatory band at *D6Mit150* regulates at least 1,560 transcripts. Numerous candidate genes lie in the flanking intervals from *D6Mit10* to *D6Mit254*. Several sources of evidence can be used to identify candidate genes: (**a**) the SNP density and presence of missense SNPs, (**b**) the *cis* regulation of transcript abundance and (**c**) high expression of transcripts in brain and other neural tissues from the GNF Expression Atlas data track superimposed on the University of California Santa Cruz genome browser.
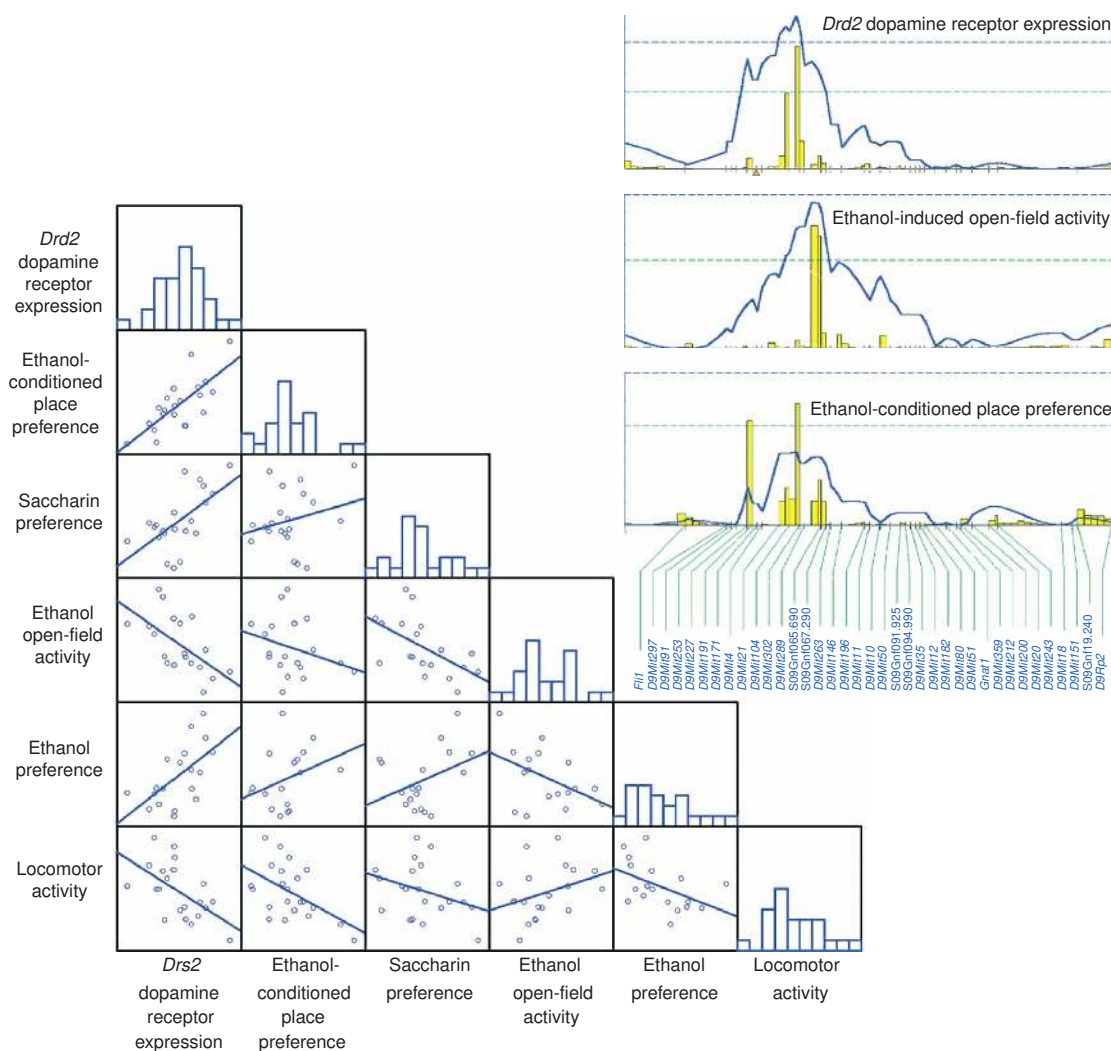
Figure 5 Genetic correlation of *Drd2* expression with several behavioral phenotypes in WebQTL's BXD Published Phenotypes database. In this analysis, trait means from each of the BXD strains for which both microarray data and behavioral data exist are correlated. The correlated traits are ethanol conditioned place preference (record number 10542), saccharin preference (10542), ethanol-induced open-field activity (10076), ethanol preference (10477) and locomotor activity (10485). Inset, Chromosome 9 interval maps for *Drd2* expression, ethanol-induced open-field activity and ethanol-induced conditioned place preference. The allelic effect is in the opposite direction for open-field activity, a trait that is negatively correlated with *Drd2* expression. This analysis was done using MAS 5.0–normalized expression data.

(*Zfp9*, *Zfp422* and *Zfp239*). Of these, *Foxj2* has an intron SNP, and *Zfp9* has three SNPs in the 3′ untranslated region (UTR), one SNP in the 5′ UTR and several intron SNPs between the progenitors. There are 38 missense SNPs in the region flanking this marker, which may disrupt function of *Mbd4*, *D6Wsu116e*, *Cacna1c*, *Bcl2l13*, *Bid*, *Usp18* and *A2m*. Additionally, there are 64 SNPs in the 3′ UTR and two in the 5′ UTR.

QTLs can be caused by polymorphisms that affect gene expression rather than protein sequence. Polymorphisms controlling variation of large groups of transcripts may therefore be manifested as *cis*-regulatory QTLs that occur at the location of a *trans*-regulatory band. Several *cis*-regulatory QTLs are found at the *D6Mit150* region (**Fig. 4**), including *Camk1* (160882_at), *Rho* (96567_at), *A2m* (104486_at), *Phc1* (100992_at), *Slc6a1* (161059_at), *Itpr1* (93895_s_at) and *Apobec1* (98398_s_at).

The *Mtap2* locus (on chromosome 1 at 67.928 Mb) modulates several hundred transcripts, including several motor proteins and neurotransmitter receptors. This marker, which is a highly polymorphic gene, is a compelling candidate for regulation of these transcript abundances. *Mtap2* is a modestly *cis*-regulated transcript, and expression is correlated with the expression of several motor proteins (*Kif1b*, *Kif5a* and *Kif5c*), clathrin-coated vesicle proteins (*Ap1g1*, *Syt1* and *Tgoln2*) and neurotransmitter receptors (*Gabra1*, *Gabra3*, *Gria1*, *Gria3* and *Grik2*). These are just a subset of the over-represented Gene Ontology categories among *Mtap2* correlates. *Mtap2* contains at least seven missense polymorphisms between strains C57BL/6J and DBA/2J. Examination of multiple types of converging evidence indicates that these polymorphisms potentially have a causal role in regulation of other transcripts mapping to the *Mtap2* locus. Similar analyses can be done for the other *trans*-regulatory QTL bands and for individual transcripts.

Tissue specificity of expression regulation

Bystrykh and colleagues[14] used essentially identical methods to study the genetic modulation of transcriptional activity in flow-sorted
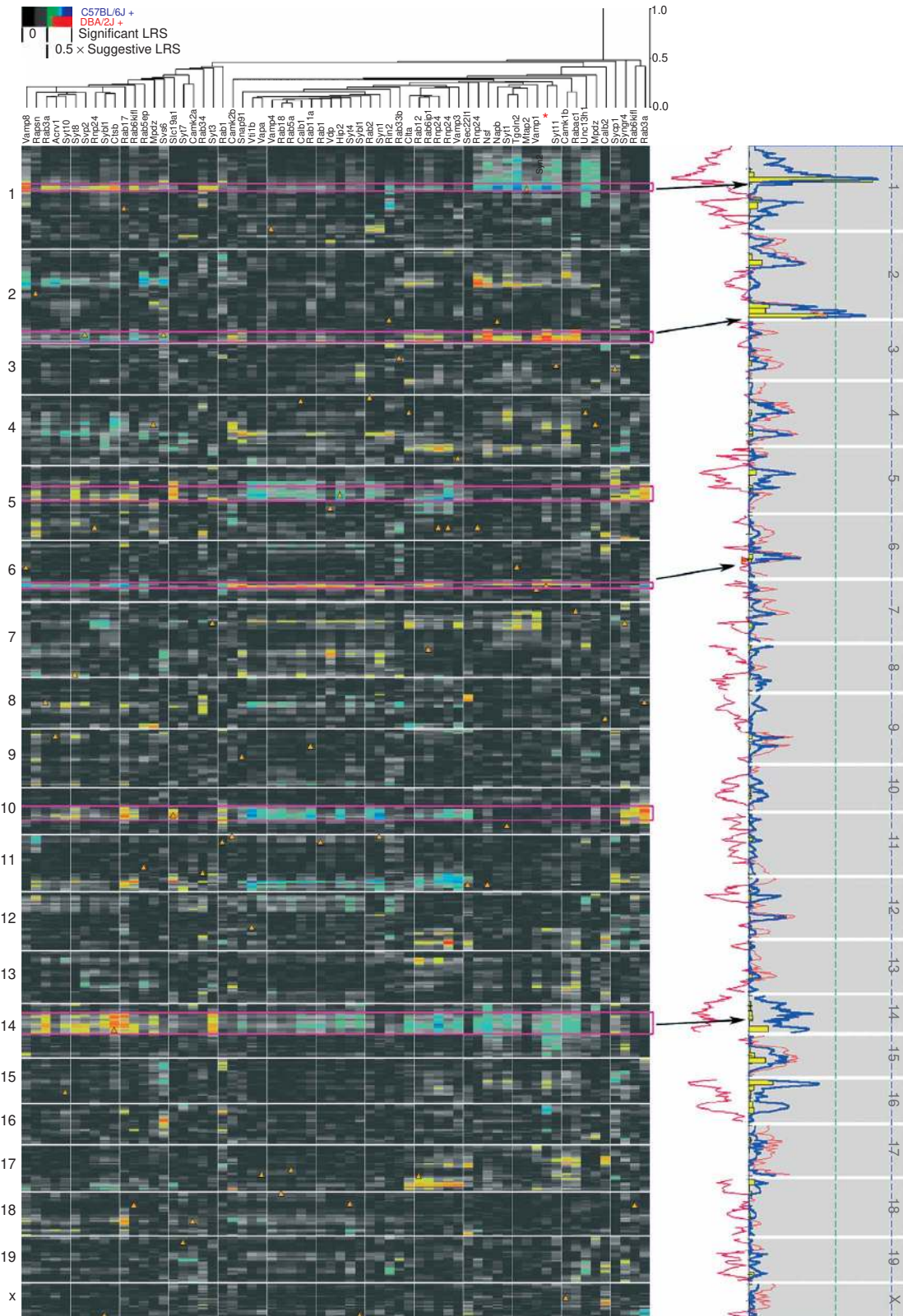
Figure 6 Cluster map showing the polygenic and pleiotropic regulation of the synaptic vesicle cycling mechanisms. Transcripts related to synaptic vesicle cycling are clustered based on their genetic correlations, and regulatory loci are mapped along the *y* axis of the main plot. Cool hues represent LRS for elevated transcription in mice with C57BL/6J genotypes at a given locus, and warm hues represent LRS for elevated transcription in mice with the DBA/2J allele. The single QTL scan on the right side of the plot is for *Syn2* expression (red asterisk on the dendrogram). Several regulatory loci seem to control the expression of multiple synaptic vesicle components in parallel. These loci include some of the key *trans*-regulatory QTL bands (*e.g.*, *Mtap2* on chromosome 1).

HSCs of BXD strains. Although most global regulators of transcription are tissue-specific, a considerable number of regulatory loci modulate the same transcript in both the brain and HSCs. Of the top 76 *cis*-regulatory QTLs that we detected in the brain, 35 were also significant in the RMA-normalized HSC data set (**Supplementary Table 4** online), suggestive of a common regulatory process rather than a neural-specific one. This suggests that expression profiling of lymphocytes could provide insight into genetic variation in gene expression in the central nervous system that is relevant to human neurological disease. Seven others were merely suggestive in the HSC data, and the remainder had either non-significant local maxima in the region (19 probe sets) or no *cis*-acting QTLs (22 probe sets). These latter 22 transcripts have *cis*-acting modulation that is tissue-specific. *Trans*-regulatory QTLs were also tissue-specific. Four *trans*-regulatory QTLs detected in the central nervous system were replicated in the HSC analysis, but unique *trans*-regulatory QTLs were identified in both tissues for at least four other transcripts.

Epistatic control of variation in gene expression

Genetic variation in expression is often produced by combinations of QTLs that have independent effects or that interact to produce epistatic effects[3]. Expression of *Grin2b*, encoding an NMDA receptor subunit, is modulated by loci on chromosomes 6 and 8 (ref. 24). These loci account for ~50% of the variance in expression. A reanalysis of *Grin2b* using a model that includes a two-locus epistatic interaction also fits the data well. This model highlights a locus on distal chromosome 2 near *D2Mit148* that has no additive effect but interacts strongly with the *cis*-regulatory QTL on distal chromosome 6. These two loci with their interaction explain 89% of the total genetic variance, and the interaction is significant at $P < 0.01$. The main caveat of systematically searching for epistatic interactions for all transcripts is the risk of overfitting the model when using a small number of strains. The 32 genotypes in this particular case are well-distributed among the four digenic states, but to improve long-term utility of the BXD set for this purpose, we generated additional lines that increase the possible sample size to ~80 strains[7].

Association of variation in expression with behavior

The utility of RI lines extends beyond their use as a mapping panel. Unlike conventional F_2 and backcross progeny, RI mapping panels can be reproduced indefinitely, making it practical to extend studies across treatments, ages and environments[25,26]. The BXD RI lines have already been used to analyze several hundred neurological and behavioral phenotypes, which we have assembled in WebQTL. Strong statistical associations are often detected between transcripts and neuroanatomical[27,28] or behavioral traits. For example, expression of the D2 dopamine receptor (*Drd2*, 97776_at) is correlated with midbrain iron levels in female mice[29] ($r = 0.70$, $P < 0.003$), ethanol-induced conditioned place preference[30] ($r = 0.52$, $P < 0.009$) and other phenotypes (**Fig. 5**)[31–34]. One of several loci that modulate *Drd2* mRNA levels is located close to *Drd2* itself. This chromosome 9 locus near *D9Mit302* is important in ethanol-induced open-field activity[32]. Although the *Drd2* transcript itself is not polymorphic[35], several SNPs have been identified in its promoter. Remapping conditioned place preference using WebQTL's Published Phenotypes database and high-precision genetic map identifies several suggestive QTLs. A regulatory peak on chromosome 9 indicates that the polymorphism regulating *Drd2* may also influence this behavioral trait (**Fig. 5**) or that multiple linked polymorphisms in this region simultaneously affect the phenotypes in parallel.

Associative networks of transcriptional control

Associative networks can be rapidly assembled from the covariance matrix of molecular, cellular and behavioral traits and their shared upstream regulatory loci. Gene-to-gene correlations have a low rate of false positive associations across the overall data set. For correlations of 0.58 and above, an FDR[36] of 1% is obtained with 35 strains, even with 77 million implicit tests. Correlations of strain means for transcripts with moderate to high expression levels primarily reflect shared genetic rather than environmental or technical effects on gene expression, because multiple individuals were used to create a within-strain phenotype mean. Genetically correlated traits, whether transcripts, neuroanatomical traits or behaviors, by definition share common QTLs.

We mapped the joint modulation of synaptic vesicle–related transcripts using associative network tools in WebQTL. These transcripts include synaptotagmins, synapsins, synaptogyrins, synaptic vesicle proteins, vesicle-associated membrane proteins, rapsins, trans-Golgi network proteins, Rab proteins, Cam kinases and multiple pdz domain proteins (**Fig. 6**). Most of the *trans*-regulatory QTL bands observed on the transcriptome map modulate transcription of genes in this functional category, including the *Mtap2* locus on chromosome 1. Some of the correlated transcripts are also *cis*-regulated at these loci, making them candidate genetic modifiers of synapse-related transcription. These transcripts include *Mtap2* on chromosome 1, *Svs6* and *Svp2* on chromosome 2, *Hip2* on chromosome 5, *Syn2* on chromosome 6, *Slc19a1* on chromosome 10 and *Ctsb* on chromosome 14. *Mtap2* mRNA targets the synapse[37] and contains numerous missense polymorphisms, making it a prime candidate for the genetic modulation of synaptic mRNA. The control of synaptic transcripts by the *trans*-acting bands indicates that variation in this pathway is a key genetic difference in brain function among BXD RI strains. These broad differences in transcript abundance across strains could have numerous pleiotropic effects on characteristics from synaptic efficacy to behavior.

Cliques of transcripts and behavioral phenotypes

Networks of biological traits are widely thought to be scale-free[38], meaning that a relatively small number of transcripts and gene regulatory vertices are highly connected hubs whereas others interact selectively with only a few transcripts. New algorithms[39] were applied to extract groups of highly interconnected transcripts (cliques) from genetic correlation matrices containing millions of expression level correlations. We identified *Lin7c* as one of the most highly connected transcripts in the brain. One clique consisted of 17 highly correlated transcript abundances for several mRNA spliceosome-related proteins. This clique overlaps in composition with more than 1,700 other cliques. Using a near-clique algorithm called paraclique, we combined many of these cliques into one large group of 193 transcripts by extracting highly (but not perfectly) interconnected sets of transcripts. Among the members of the largest paraclique is *Cask*, which encodes a synaptic protein that physically interacts with *Lin7c*. Multiple QTL mapping analysis showed that many of this paracliques' members are regulated by loci near *D6Mit150* and *D12Mit146*, following a general pattern in which each paraclique is regulated by combinations of *trans*-acting bands and other loci. Two clique members are located within this QTL: B-cell receptor associated protein (*Bcap29*, 160876_at) and myelin transcription factor 1-like (*Myt1l*, 96496_g_at). These clique members are high-priority candidate genes for modulation of the massive *Lin7c* clique. Expression of the *Lin7c* clique members correlates with both midbrain iron levels[28] and several locomotor behavior measures. Notably, one of the clique members, *Strn3*, is a striatin family member that is also associated with

locomotor impairment[40]. Further, *Cask*, *Lin7c*[41] and locomotor activity are associated with the expression of serotonin receptor *Htr2c*. These two proteins form a complex with Mint1 that has a key role in synapse function[41]. The detection of the relationships of many new genes to a behavior allows substantial expansion of the molecular pathway underlying this phenotype.

DISCUSSION

Naturally occurring genetic polymorphisms alter gene expression in the central nervous system in a massively correlated fashion. A very small subset of polymorphisms contributes to variation in a large set of transcripts. For example, a locus near *D6Mit150* modulates the expression of at least 1,650 transcripts (**Fig. 4**) and is a good target for interval reduction and candidate gene analysis. The impact of these polymorphisms is often tightly coupled with variation in receptor density, neuron number, neuronal excitability and, ultimately, behavior (**Fig. 5**). The accumulation of additional phenotypes will elucidate the functional importance of this massive genetic covariation.

Previous genome-wide studies of regulation of gene expression identified loci that coregulate many transcripts[3,12], but these loci were assumed to act independently. We showed that sets of loci cooperatively influence large sets of phenotypes, including many transcripts that underlie synaptic function (**Fig. 6**). Because of the multi- and polygenic nature of brain transcription control, detection of mendelian loci will be infrequent. Furthermore, the high degree of covariance among transcriptional phenotypes is a substantial challenge to most attempts to control the family-wise error rate. Despite these challenges, we detected statistically significant QTLs for transcript abundance that segregate among BXD RI lines (**Fig. 1** and **Supplementary Table 2** online). An examination of genome-wide QTL frequency identifies the locations of key genetic regulatory loci (**Figs. 3** and **4**). The cluster map (**Fig. 6**) emphasizes extensive coregulation of transcripts. Sets of modulatory loci seem to be tissue-specific[14] and probably vary with experimental and environmental perturbation.

We have begun functional annotation of these genetically coregulated systems by integrating transcriptome-QTL and genetic correlation of gene expression in a panel of RI strains. A deceptively simple concept is at the heart of this work: the use of a stable GRP to study a diverse range of phenotypes and phenomena. GRPs provide an efficient analytical approach to synthesize growing collections of biological data across many levels of organization and can be extended to virtually any tissue or cell type, from single cells to the complex mammalian brain. We extended this resource to a total of 80 BXD lines; therefore, power and precision will improve substantially[7].

Transcriptome QTL analysis of numerous other tissues is progressing rapidly and will allow the research community to detect the shared and unique components of tissue-specific transcription-regulatory machinery (**Supplementary Table 4** online). The identification of tissue-specific loci uncovers mechanisms that underlie development and maintenance of tissue differentiation. It is also now practical to identify loci responsible for wide-ranging changes in gene expression triggered by exposure to pharmacological agents, pathogens, environmental stressors and natural processes of development and aging. Refinement of new technologies for high-throughput phenotypic assays including proteomics[42] and sequencing will enable evaluation of the micro- to macro-scale genetic effects across tissues and environments.

A multitude of hypotheses can be developed or tested using the transcriptome and systems-level phenotype data incorporated into WebQTL[24]. Query-specific multiple testing adjustment can be used to estimate significance thresholds. Large inductive queries aimed at producing new hypotheses for costly experimental follow-up require

stringent thresholds. In contrast, more focused or confirmatory queries, such as those aimed at identifying behavioral correlates of *Drd2* expression, require less stringent error control because of the large body of existing knowledge driving the research question. Although we focused on a small set of reliable QTL results (**Supplementary Table 2** online), all trait data are available on WebQTL, for users who have existing information about specific regulatory relations. The 88 conservatively chosen QTLs are fewer than the expected number of false positives based on a null distribution of 12,422 hypotheses. But the number of implicit hypotheses is actually much smaller than the number of probe sets on the array, for two reasons. First, high covariance among transcript phenotypes leads to massive dependence of the statistical tests. Second, we considered only subsets of traits with higher heritability. The incorporation of existing knowledge when defining a hypothesis set is a powerful approach for reduction of false positives. Ultimately, users of this resource must consider the appropriate error thresholds based on the relative practical consequences of false positive or false negative results.

A synergistic combination of positional precision and comprehensive bioinformatics resources can be exploited to identify causative polymorphisms for gene expression covariation[43]. This synergy is particularly strong for the BXD RI lines because the progenitor strains have been almost completely sequenced, simultaneously allowing interval reduction using fine-grained haplotype structure[44] and evaluation of functional consequences of precise polymorphisms on genes. Gene selection and identification requires multiple sources of evidence that converge on a subset of positional QTL candidates (**Fig. 4**). Criteria used to evaluate candidates include identification of missense polymorphisms, sequence conservation, level of gene expression in the relevant tissue and stage, and evidence that a candidate gene is itself under *cis* regulation. The evidence is even stronger when plausible biological models already predict causal relations between candidates and target transcripts. In the case of DNA-binding transcription factors, evidence of candidacy may come from a common promoter motif found among target genes. Literature mining and gene ontology[21] analysis of *trans*-regulated genes can also be used to identify candidates. Ultimately, a small set of genes within the region can be selected for functional and molecular studies.

Individual differences in brain and behavior are produced by genetic and environmental effects that often act through the modulation of mRNA transcription. The shared mediators of gene expression simultaneously alter hundreds of transcript levels and physiological, morphological and behavioral phenotypes. The unique properties of GRPs allow this multiscalar biological data to be harnessed for rapid refinement and identification of key polymorphic genes. The causal relations provided by transcription QTL mapping greatly facilitate specification of genetic regulatory networks. By integrating data from base pair to behavior in a single reference population, testable networks of the effects of genetic and environmental variation across all levels of biological scale can be developed. It is now possible to define multifactorial genetic and environmental influences on transcriptional modules and systems-level phenotypes as they change during development, aging and disease.

METHODS

BXD RI mice. We measured steady-state transcript abundance in a panel of BXD RI strains, both parental strains and the C57BL/6J × DBA/2J F$_1$ hybrid (a total of 35 isogenic lines). To generate the BXD RI set, we crossed progenitor strains C57BL/6J and DBA/2J strains and mated them to their siblings for more than 36 generations. This resulted in a panel of inbred strains with fixed genotypes at each locus, with parental C57BL/6J and DBA/2J alleles segregating

among the strains. There are now 35 commercially available BXD RI lines and 45 new lines at the University of Tennessee Health Science Center[7].

Genotypes database. Together with our colleagues, we genotyped more than 1,500 marker loci in this RI set, resulting in a very dense error-checked map consisting of 779 nonredundant loci[45]. The mean precision of RI mapping resources is currently ~4 Mb for mendelian traits.

Phenotypes database. We assembled a comprehensive and complementary database that integrates published phenotypes for BXD and other RI strains. This database contains data types as diverse as dentate granule cells numbers, alcohol preference, maze learning and open-field activity levels. These classical phenotypes can thus be readily compared with variation in abundance of all transcripts using WebQTL. This interactive web system for complex trait analysis allows users to adjust analysis parameters and to analyze the array data presented here with respect to their own phenotypic assays.

Array annotation. Probe positions in WebQTL are determined by systematic BLAT analysis of the concatenation of all 16 perfect-match 25-nt probe sequences (corrected for probe overlap) against the current version of the public mouse assembly at (our results are based on the October 2003 freeze). The position and identity of many thousands of the transcripts that are targeted by the Affymetrix U74Av2 probe sets were manually error-checked and curated over a 3-year period resulting in a greatly improved annotation for this particular array platform.

Screening for SNPs in probes. Only a small fraction of the *cis*-regulatory QTLs might be attributable to known SNPs in the probe sequence. In an analysis of all 1.2 million known SNPs segregating between C57BL/6J and DBA/2J, we found 651 SNPs in 1,223 of the Affymetrix U74Av2 perfect-match probe sequences using the Celera SNP database (1 July 2003). Roughly 1 in 10 of these is sufficient to affect probe hybridization differences between the strains enough to create an artefactual QTL (**Supplementary Table 4** online).

Tissue processing and gene expression. Most expression data are strain averages based on three microarrays (U74Av2). Each individual array experiment involved a pool of brain tissue (forebrain plus the midbrain, but without the olfactory bulb, retina or neurohypophysis) that was taken from three adult mice, usually of the same age. We used 100 arrays: 74 female pools and 26 male pools. Mice ranged in age from 56 to 441 days and typically included one pool at 8 weeks, one pool at ~20 weeks and one pool at ~1 year. Only a small fraction of sex differences and age differences were identified in analyses of balanced or fully saturated (representation of all combinations of sex × strain × age) subsets of these data. Each data table in WebQTL has a link to detailed metadata on the experiment, analysis method and samples comprising the data set.

Array normalization. We transformed data using several common array normalization methods, including RMA, MAS 5.0, dChip (PM, PMMM) and PDNN. Expression data for each normalization method are available at WebQTL. Our mapping results here used the relatively conservative RMA[46] method. All normalizations were done using default analysis parameters. For MAS 5.0–normalized data, additional processing occurred. We generated probe set data using MAS 5.0, obtained the $\log_2$ of each probe set and standardized using Z scores. We doubled the Z scores and added 8 to produce a set of Z scores with a mean of 8, a variance of 4 and a standard deviation of 2. The advantage of this modified Z score is that a twofold difference in expression level corresponds approximately to a 1-unit difference. Expression levels below 5 are usually close to background noise levels. All DAT, TXT, RPT, CEL and CHP files for the 100 arrays in this report are available in WebQTL.

The treatment of probe set–level data is an active area of development, and numerous methods of using this data have been proposed. But few methods have considered the differences between individual probes and have often treated each probe as an equally valid measure of the same transcript. In many cases, probes overlap highly in sequence, and those probes often detect highly correlated expression[43]. Several probes match (using BLAT) multiple regions of the genome and may therefore bind to several mRNAs. Furthermore, as noted above, SNPs are present in a small number of probes, rendering affinity higher for one allele than another. Annealing temperatures vary between the probes,

and the particular exon binding of the probes varies, such that individual probes may represent different splice variants. Physical characteristics, exon binding and genome location of probes along with a direct link to the BLAT alignment program at the University of California Santa Cruz's Genome Browser are available at WebQTL in the probe information tables.

Variance partitioning. We estimated between- and within-strain variance components using SAS 9.0. We calculated the genetic variance accounted for by strain (a measure related to the heritability) from the ratio of between-strain variance to total phenotypic variance (the between-strain intraclass correlation). This measure estimates genetic variance relative to environmental and technical variability. Environmental variance is minimized by using pooled samples on each array. We obtained the standard error of the intraclass correlation using a formula for obtaining the variance of intraclass correlations in the presence of data imbalance and single observations. We obtained adjusted heritabilities using a formula to adjust for the overestimation of the additive effect in inbred strains.

QTL mapping. We carried out linkage mapping for 12,422 transcript expression traits. We excluded parental and F_1 lines from the mapping analysis. Mapping was done using strain averages of probe set expression levels obtained using RMA. QTL mapping was done using a custom program, QTL Reaper, that carries out simple regression implemented in Python and C. Performance-critical code was implemented in C and compiled as a Python module that is also used by WebQTL. We estimated genome-wise empirical *P* values by permuting trait data for each transcript randomly between 1,000 and 1,000,000 times[47]. We obtained confidence intervals by bootstrap analysis[48,49]. We estimated a point-wise FDR, the q value, for the set of transcripts declared significant at each transcript specific QTL *P* value. This approach is used for estimating the error rate among the large set of hypotheses tested across the microarray. Two separate permutation analyses applied to the transcriptome map show that the *trans*-regulatory QTL bands are not an artefact of the genotypic structure of the RI strain panel but are due to the correlation and, therefore, coregulation by one or more closely linked regulatory QTLs. The *trans*-acting bands disappear entirely when the panel of markers is left intact but the transcript expression levels are each individually permuted, indicating that the correlation and potential coregulation is the explanation for *trans* regulation. Permutation of the data by randomly assigning entire chips to the genotypes results in maintained structure, but not location, of the *trans*-acting bands, showing that the position of *trans*-regulatory QTL bands is not a statistical consequence of location-specific bias in marker strain distribution patterns. We carried out pair-wise QTL scanning for epistasis for a small number of selected transcripts using R/qtl[50]. The multiple QTL models were significant at $P < 0.01$ based on whole-genome permutations of a pair-wise scan. Sample sizes for each chromosome 2–chromosome 6 configuration in the *Grin2b* analysis are B/D = 7, D/B = 8, B/D = 10 and B/B = 7 (where D is DBA/2J and B is C57BL/6J).

Genetic correlation analysis. We computed Pearson product-moment correlations of strain means for each pair of probe sets on the array. Both Spearman's rank correlations and Pearson product-moment correlations can be computed using WebQTL. FDR estimation for the entire gene-gene correlation matrix was implemented in PERL.

Clique extraction. We carried out clique analysis on data transformed by RMA and MAS 5.0. We extracted cliques using algorithms and hardware developed at University of Tennessee Knoxville. In this analysis, an edge-weighted graph is constructed from the entire correlation matrix. The vertices of this graph represent genes labeled with probe set identifications. Each pair of vertices is connected by an edge whose weight is taken from the correlation matrix. A high-pass filter is used to eliminate any edge whose weight $|r|$ is less than 0.85. The remaining edges are then unweighted. Cliques are identified through a transformation to the complementary dual vertex cover problem and the use of highly parallel algorithms based on the notion of fixed-parameter tractability.

Animal care. All procedures involving mouse tissue were approved by the Institutional Animal Care and Use Committee at the University of Tennessee Health Science Center.

URLs. WebQTL, which includes all gene expression data, sample preparation information, the Published Phenotypes database and a suite of interactive tools for analysis of recombinant inbred mouse phenotypes, is available at http://www.webqtl.org/. It is the first component of the Gene Network (http://www.genenetwork.org/). Additional information including detailed experimental procedures and a wealth of RI phenotypic data is available at http://www.nervenet.org/ and http://www.mbl.org/. Gene Ontology Tree Machine is available at http://genereg.ornl.gov/gotm/ and can be invoked directly from WebQTL. The public mouse genome assembly is available at http://genome.ucsc.edu/cgi-bin/hgBlat?command=start/.

Note: Supplementary information is available on the Nature Genetics website.

1. Cowles, C.R., Hirschorn, J.N., Altshuler, D. & Lander, E.S. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32**, 432–437 (2002).
2. Jansen, R. & Nap, J.P. Genetical genomics: the added value from segregation. *Trends Genet.* **17**, 388–391 (2001).
3. Brem, R., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 725–755 (2002).
4. Bading, H., Ginty, D.D. & Greenberg, M.E. Regulation of gene expression in hippocampal neurons by distinct calcium signaling pathways. *Science* **260**, 181–186 (1993).
5. Paigen, K. & Eppig, J. A mouse phenome project. *Mamm. Genome* **111**, 715–717 (2000).
6. Threadgill, D.W., Hunter, K.W. & Williams, R.W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mamm. Genome* **13**, 175–178 (2002).
7. Peirce, J.L. *et al.* A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* **5**, 7 (2004).
8. Li, X. *et al.* High-resolution genetic mapping of the saccharin preference locus (Sac) and the putative sweet taste receptor (T1R1) gene (Gpr70) to mouse distal Chromosome 4. *Mamm. Genome* **12**, 13–16 (2001).
9. Mogil, J.S. *et al.* The melanocortin-1 receptor gene mediates female-specific mechanisms of analgesia in mice and humans. *Proc. Natl. Acad. Sci. USA* **100**, 4867–4872 (2003).
10. Shirley, R.L. *et al. Mpdz* is a quantitative trait gene for drug withdrawal seizures. *Nat. Neurosci.* **7**, 699–700 (2004).
11. Taylor, B.A. *et al.* Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. *Mamm. Genome* **10**, 335–348 (1999).
12. Flint, J. Analysis of quantitative trait loci that influence animal behavior. *J. Neurobiol.* **54**, 46–77 (2003).
13. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
14. Bystrykh, L. *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* advance online publication, 13 February 2005 (doi:10.1038/ng1497).
15. Belknap, J.K. Effect of within-strain sample size on QTL detection and mapping using recombinant inbred mouse strains. *Behav. Genet.* **28**, 29–38 (1998).
16. Wang, J., Williams, R.W. & Manly, K.F. WebQTL: Web-based complex trait analysis. *Neuroinformatics* **1**, 299–308 (2003).
17. Rikke, B.A. & Johnson, T.E. Towards the cloning of genes underlying murine QTLs. *Mamm. Genome* **9**, 963–968 (1998).
18. Visscher, P.M. Speed congenics: accelerated genome recovery using genetic markers. *Genet. Res.* **74**, 81–85 (1999).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
20. Manly, K.F., Nettleton, D. & Hwang, J.T. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* **14**, 997–1001 (2004).
21. Storey, J.D., Taylor, J.E. & Siegmund, D. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B* **66**, 187–205 (2004).
22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
23. Zhang, B., Schmoyer, D., Kirov, S. & Snoddy, J. GOTree Machine (GOTM): a web-based platform for interpreting interesting sets of genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**, 16 (2004).
24. Chesler, E.J., Lu, L., Wang, J., Williams, R.W. & Manly, K.F. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat. Neurosci.* **7**, 485–486 (2004).
25. Chesler, E.J. *et al.* Genetic correlates of gene expression in recombinant inbred strains: a relational model system to explore neurobehavioral phenotypes. *Neuroinformatics* **1**, 343–357 (2003).
26. Churchill, G.A. *et al.* The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36**, 1133–1137 (2004).
27. Lu, L., Airey, D.C. & Williams, R.W. Complex trait analysis of the hippocampus: mapping and biometric analysis of two novel gene loci with specific effects on hippocampal structure in mice. *J. Neurosci.* **21**, 3503–3514 (2001).
28. Peirce, J.L., Chesler, E.J., Williams, R.W. & Lu, L. Genetic architecture of the mouse hippocampus: identification of gene loci with selective regional effects. *Genes Brain Behav.* **2**, 238–252 (2003).
29. Jones, B.C. *et al.* Quantitative genetic analysis of ventral midbrain and liver iron in BXD recombinant inbred mice. *Nutr. Neurosci.* **6**, 369–377 (2003).
30. Cunningham, C.L. Localization of genes influencing ethanol-induced conditioned place preference and locomotor activity in BXD recombinant inbred mice. *Psychopharmacology* **120**, 28–24 (1995).
31. Risinger, F.O. & Cunningham, C.L. Ethanol-induced conditioned taste aversion in BXD recombinant inbred mice. *Alcohol. Clin. Exp. Res.* **22**, 1234–1244 (1998).
32. Crabbe, J.C., Kosobud, A., Young, E.R. & Janowsky, J.S. Polygenic and single-gene determination of responses to ethanol in BXD/Ty recombinant inbred mouse strains. *Neurobehav. Toxicol. Teratol.* **5**, 181–187 (1983).
33. Phillips, T.J., Crabbe, J.C., Metten, P. & Belknap, J.K. Localization of genes affecting alcohol drinking in mice. *Alcohol. Clin. Exp. Res.* **18**, 931–941 (1994).
34. Phillips, T.J., Huson, M., Gwiazdon, C., Burkhart-Kasch, S. & Shen, E.H. Effects of acute and repeated ethanol exposures on the locomotor activity of BXD recombinant inbred mice. *Alcohol. Clin. Exp. Res.* **19**, 269–278 (1995).
35. Hitzemann, R. *et al.* Dopamine D2 receptor binding, Drd2 expression and the number of dopamine neurons in the BXD recombinant inbred series: genetic relationships to alcohol and other drug associated phenotypes. *Alcohol. Clin. Exp. Res.* **27**, 1–11 (2003).
36. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
37. Blichenberg, A. *et al.* Identification of a cis-acting dendritic targeting element in MAP2 mRNAs. *J. Neurosci.* **19**, 8818–8829 (1999).
38. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabási, A.-L. The large scale organization of metabolic networks. *Nature* **407**, 651–653 (2000).
39. Baldwin, N.E. *et al.* Computational, integrative and comparative methods for the elucidation of genetic co-expression networks. *J. Biomed. Biotechnol.* (in the press).
40. Bartoli, M. Down-regulation of striatin, a neuronal calmodulin-binding protein, impairs rat locomotor activity. *J. Neurobiol.* **40**, 234–243 (1999).
41. Becamel, C. *et al.* Synaptic multiprotein complexes associated with 5-HT(2C) receptors: a proteomic approach. *EMBO J.* **21**, 2332–2342 (2002).
42. Klose, J. *et al.* Genetic analysis of the mouse brain proteome. *Nat. Genet.* **30**, 385–393 (2002).
43. Chesler, E.J. & Williams, R.W. Brain gene expression: genomics and genetics. *Int. Rev. Neurobiol.* **60**, 59–95 (2004).
44. Yalcin, B. *et al.* Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. *Nat. Genet.* **36**, 1197–1202 (2004).
45. Williams, R.W., Gu, J., Qi, S. & Lu, L. The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biol.* **2**, RESEARCH0046 (2002).
46. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
47. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971 (1994).
48. Visscher, P.M., Thompson, R. & Haley, C.S. Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020 (1996).
49. Hutson, A.D. Bootstrap smoothing strategies based on uniform spacings with practical applications. Technical Report. (Division of Biostatistics, University at Buffalo, Buffalo, New York, 2002).
50. Broman, K., Wu, H., Sen, S. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics,* **19**, 889–890 (2003).