Copy number variation influences gene expression and metabolic traits in mice

Luz D. Orozco1,*, Shawn J. Cokus2, Anatole Ghazalpour3, Leslie Ingram-Drake1, Susanna Wang1, Atila van Nas1, Nam Che3, Jesus A. Araujo3, Matteo Pellegrini2 and Aldons J. Lusis1,3

1Department of Human Genetics, UCLA, Los Angeles, CA, USA, 2Department of Molecular, Cell and Developmental Biology, UCLA, Los Angeles, CA, USA and 3Division of Cardiology, Department of Medicine, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA

Received March 19, 2009; Revised June 17, 2009; Accepted July 28, 2009

Copy number variants (CNVs) are genomic segments which are duplicated or deleted among different individuals. CNVs have been implicated in both Mendelian and complex traits, including immune and behavioral disorders, but the study of the mechanisms by which CNVs influence gene expression and clinical phenotypes in humans is complicated by the limited access to tissues and by population heterogeneity. We now report studies of the effect of 19 CNVs on gene expression and metabolic traits in a mouse intercross between strains C57BL/6J and C3H/HeJ. We found that 83% of genes predicted to occur within CNVs were differentially expressed. The expression of most CNV genes was correlated with copy number, but we also observed evidence that gene expression was altered in genes flanking CNVs, suggesting that CNVs may contain regulatory elements for these genes. Several CNVs mapped to hotspots, genomic regions influencing expression of tens or hundreds of genes. Several metabolic traits including cholesterol, triglycerides, glucose and body weight mapped to three CNVs in the genome, in mouse chromosomes 1, 4 and 17. Predicted CNV genes, such as Itlna, Defcr-1, Trim12 and Trim34 were highly correlated with these traits. Our results suggest that CNVs have a significant impact on gene expression and that CNVs may be playing a role in the mechanisms underlying metabolic traits in mice.

INTRODUCTION

Copy number variants (CNVs) are DNA segments with a variable number of repeats among individuals, ranging from kilobases to several megabases in length. CNVs are an important source of genetic variation in diverse human populations (1,2), as well as in primates (3,4) and rodents (5,6). CNVs can influence gene expression (7), presumably by altering gene dosage, through disruption or duplication of CNV regions containing genes. In humans, they are associated with a number of Mendelian and complex genetic disorders, including autoimmune disease (8), HIV infection (9,10) and autism (11). The mechanisms by which CNVs contribute to disease in humans have been difficult to study, in part due the difficulty in obtaining tissue samples and population heterogeneity. However, the presence of CNVs in mouse inbred strains, as well as the ability to manipulate the mouse genome to map gene expression and clinical traits using crosses, makes the mouse an ideal model to dissect the biological significance of CNVs. Analyses of CNVs in mouse genomes have demonstrated significant variation among mouse inbred strains (5,12) as well as non-random recurrent CNVs among members of the same inbred strain (13). Moreover, CNVs identified between mice of different inbred strains are of similar size and magnitude as those identified among different human populations (1,5,12), suggesting that the mouse could serve as a model to study the biological significance of CNVs.

In order to establish whether the mouse can be used as a model to study the impact of copy number variation, we investigated the effect of CNVs on gene expression phenotypes using a panel of CNVs previously identified in 20 mouse inbred strains (5). We asked whether CNVs influenced gene expression or clinical traits by using this set of CNVs, in conjunction with genome-wide gene expression and metabolic trait data from two independent mouse crosses between strains C57BL/6J and C3H/HeJ. We and others have shown

*To whom correspondence should be addressed. Tel: +1 3108251359; Fax: +1 3107947345; Email: LDORozco@mednet.ucla.edu

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
that the genetics of gene expression can be used as a link between DNA variation and phenotypic traits to prioritize candidate genes and to identify causal relationships between chromosomal regions and clinical phenotypes (14–16). Here we show that mouse CNVs resulted in altered gene expression in the genes mapping to CNVs, which was highly correlated with copy number. We also observed an effect in genes flanking CNVs, suggesting that CNVs can influence gene expression through disruption of regulatory sequences. Our results also show that expression QTL (eQTL) hotspots mapped to CNVs, suggesting that regulatory elements present in CNVs and/or eQTL mapping to CNVs may be influencing the expression of hundreds of genes in \textit{trans}. Furthermore, we found that a variety of metabolic traits, including body weight, cholesterol and glucose levels, mapped to a subset of the CNVs. Our results indicate that mouse inbred strains can be used to examine the mechanisms by which CNVs influence complex traits.

**RESULTS**

To assess the impact of CNVs on gene expression, we examined 19 CNVs variable between the mouse strains C57BL/6J (B6) and C3H/HeJ (C3H). These CNVs are distributed among 11 chromosomes and range in size from 47 kb to 1.9 Mb in length, with a median size of 195 kb, and contain a total 54 genes. Sixteen of 19 CNVs contain at least one gene, and 14 CNVs contain more than one gene. CNV genomic positions and array comparative genome hybridization (aCGH) ratios can be found in Supplementary Material, Table S1. These 19 CNVs represent the entire set of CNVs variable between B6 and C3H identified by Graubert \textit{et al.} (5). The range in size of the CNVs in part reflects a limitation of the aCGH platform and of the CNV calling algorithm employed. The aCGH platform contained 385,000 probes spaced \( \sim 5 \) kb apart, and probe density in different genomic locations also contributed to the resolution of the platform. In addition, the CNV detection algorithm employed by Graubert \textit{et al.} (5) required a change in intensity in at least five probes in a segment, resulting in an increased ability to detect larger CNVs.

To ensure that the CNVs reported by Graubert \textit{et al.} were also present in the parental mice used in our crosses, we validated three CNVs by qPCR on the genomic DNA from parental mice of cross 2, as well as in B6 and C3H mice recently obtained from the Jackson Laboratories (Supplementary Material, Table S2). We observed a high correlation between the published \( \log_2(\text{C3H/B6}) \) ratios and our qPCR analysis to examine the allele ratios in the genes in adipose cDNA from B6, C3H and B6xC3H F1 mice from cross 1. Each of these genes maps to CNVs with higher copy in B6, and consistent with the copy number change, we find that transcript levels in these genes show a greater B6 allele fraction in the F1 mice. These results provide further evidence that CNV mapping genes are indeed regulated in \textit{cis}.

To determine whether CNV mapping genes were indeed regulated in \textit{cis}, we performed a classic \textit{cis–trans} test in the CNV mapping genes \textit{Klrk1}, \textit{CD244} and \textit{Trim12} (Fig. 2). We carried out the \textit{cis–trans} test using semi-quantitative sequence analysis to examine the allele ratios in the genes in adipose cDNA from B6, C3H and B6xC3H F1 mice from cross 1. Each of these genes maps to CNVs with higher copy in B6, and consistent with the copy number change, we find that transcript levels in these genes show a greater B6 allele fraction in the F1 mice. These results provide further evidence that CNV mapping genes are indeed regulated in \textit{cis}.

To rule out the possibility of a recombination event between each CNV mapping gene and the nearest SNP to the gene, we examined the haplotype structure of the cross and the genotype profiles surrounding each CNV. The distribution of the distances between each CNV gene and the nearest SNP is shown in Supplementary Material, Fig. S1A, where we observed a median and mean distance of 880 kb and 1.27 Mb, respectively. We expect the degree of correlation between nearby SNPs to be very high in an F1 intercross, so that the likelihood of observing a recombination event within this range is very low. Indeed, there are blocks of highly correlated SNPs that span \( \sim 20 \) Mb in chromosome 1 (Supplementary Material, Fig. S1B), with similar results observed in other chromosomes. Furthermore, we observed an average of one recombination event per chromosome in each F2, as illustrated for chromosome 1 in Supplementary Material, Fig. S1C. No recombination break points were observed at the location of genes mapping to CNVs (Supplementary Material, Fig. S1D). One possible exception to this is for CNV 19 at the proximal end of chromosome 7, where the nearest SNP is found 14 Mb from the CNV. Although we cannot rule out the possibility of a recombination...
event between CNV 19 and the nearest SNP, the high correlation observed between nearby SNPs and the number of recombination events per chromosome suggests that this is unlikely.

We next asked whether the increase or decrease in gene expression was concordant with the direction of the change in copy number. We observed that gene expression was concordant with CNVs in 84% of the CNV genes in adipose, Figure 1.
with similar results observed in other tissues (Fig. 1C and Table 2). We used a binomial random model to determine the significance of these observations. The results of this analysis, shown in Table 2, indicated that gene expression differences were concordant with copy number in adipose tissue. In order to test the overall effect of CNVs on gene expression, we also considered all genes differentially expressed. We observed that CNVs had an effect on gene expression in adipose tissue (P = 5.97E-07), brain (P = 0.01), liver (P = 3.22E-07) and muscle (P = 1.52E-03) tissues. For CNVs carrying more than one gene, we asked how the overall expression levels were affected by each CNV using the regularized gamma function. Our results indicated that 11 out of 14 (79%) CNVs which contain more than one gene had a significant impact on gene expression levels in adipose, brain, liver and muscle tissues (Supplementary Material, Table S5). Similar results were observed when we examined the effect of CNVs on gene expression using non-parametric analysis.

To estimate the false discovery rate in these observations, we repeated our analysis 1,000 times using a randomly permuted sample in which the mouse genotypes were randomly reassigned in each sample (Table 2, Supplementary Material, Tables S4, S5 and Fig. S3). We observed that the P-values from permuted samples followed the expected uniform distribution. These results support our hypothesis that CNVs have a significant impact on gene expression levels in mice.

We next asked whether CNVs played a role in the etiology of metabolic traits using two different approaches: (i) by looking for quantitative trait loci (cQTL) near CNVs and (ii) by looking for correlations between gene expression levels and the metabolic traits. In the QTL analysis, we observed that three CNVs overlapped several cQTL for weight, cholesterol, triglycerides, glucose and insulin levels on chromosomes 1, 4 and 17 (Fig. 3A–C). Furthermore, we observed that several of the genes mapping within CNVs were significantly correlated with metabolic traits (Fig. 3D–F) and observed similar results were observed in other tissues.

In order to test a causal relationship between genetic variation, differences in gene expression and clinical traits, we used the Network Edge Orienting algorithm (NEO) (19). In essence, NEO uses structural equation modeling to test the superiority of the model where genetic variation (M) influences gene expression (A), which in turn influences a trait (B), so that M → A → B. We applied NEO to each gene expression-clinical trait pair (A and B) and the nearest SNP to the gene (M) in cross 1. The two main outputs of NEO are Local Edge Orienting (LEO) scores and Root Mean Square Error of Approximation (RMSEA). RMSEA provides a measure of the goodness-of-fit for the model under investigation and LEO provides a measure of how superior the model under investigation is, relative to other models. In general, LEO scores greater than 1 and RMSEA scores less than 0.05 suggest a causal relationship between transcript levels and a clinical trait (A → B). The results of the NEO causality test for genes that met these criteria are shown in Table 3, where we provide statistical evidence for causal relationships between CNV mapping genes and clinical traits. We found that Trim12 and Trim34 in a CNV are causal for plasma cholesterol, Ilna and Gvin1 are causal for insulin levels; Ilna and Gvin1 transcript levels are causal for plasma insulin levels and Trim12 is causal for plasma glucose levels.

To determine the reproducibility of our results, we studied the effect of CNVs on gene expression in an independent intercross between B6 and C3H (20). We determined the overlap between the genes identified in each of the two crosses using the hypergeometric distribution (Table 4) and observed
a significant overlap in adipose \( (P = 2.21 \times 10^{-9}) \), brain \( (P = 4.60 \times 10^{-8}) \) and muscle tissues \( (P = 2.29 \times 10^{-8}) \) tissues. We also observed a significant overlap in the liver when females \( (P = 2.29 \times 10^{-8}) \) and males \( (P = 2.94 \times 10^{-8}) \) were analyzed separately. Gene expression and CGH log2(C3H/B6) ratios for \emph{Itlna}, a gene mapping within a CNV in mouse chromosome 1, are shown in B6 and C3H parental strains, as well as in B6 × C3H F2 mice from the first and second crosses (Fig. 4A–C). We observed consistent differences in gene expression that were concordant with CNV in both crosses (Table 2 and Supplementary Material, Table S4). Furthermore, we determined that gene expression levels were highly correlated between the two crosses in adipose \( (r = 0.87) \), brain \( (r = 0.95) \), liver \( (r = 0.97) \) and muscle \( (r = 0.98) \) tissues in females, with similar results observed in males (Fig. 4D–F, Table 4 and Supplementary Material, Fig. S4). Overall, the results obtained in the first and second crosses supported the notion that CNV has a significant effect on gene expression and metabolic traits.

We observed that a total of 1253 eQTL mapped to CNV genomic locations, suggesting that eQTL or regulatory elements mapping to CNVs may be influencing gene expression in \emph{cis} (Fig. 5, Supplementary Material, Table S7). We illustrate this observation for a CNV in the proximal end mouse chromosome 3, which is found within the 95% confidence interval of 27 eQTL. All of these eQTL showed a LOD score greater than 4.3, and 22 of the 27 eQTL showed peak marker positions at or very near the CNV. Five of eQTL appear to be mapping in \emph{cis}, whereas the remaining genes are mapping in \emph{trans} (Fig. 5A). Since no genes mapped within this CNV (Fig. 5B), we examined the degree of conservation in the genomic sequence of this CNV using the VISTA genome browser (Fig. 5C). Previous reports have shown that highly conserved non-coding sequences represent putative regulatory elements (21),

Table 2. Concordance between CNV and gene expression in cross 1

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Sex</th>
<th>Binomial P-value</th>
<th>Permutation P-value</th>
<th>False discovery</th>
<th>Concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adipose All</td>
<td>1.14E−08</td>
<td>&lt;0.1%</td>
<td>84%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adipose Female</td>
<td>2.60E−08</td>
<td>&lt;0.1%</td>
<td>81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adipose Male</td>
<td>1.53E−07</td>
<td>&lt;0.1%</td>
<td>89%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain All</td>
<td>0.02</td>
<td>2.9%</td>
<td>80%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain Female</td>
<td>0.11</td>
<td>12.6%</td>
<td>69%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain Male</td>
<td>0.01</td>
<td>2.9%</td>
<td>85%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver All</td>
<td>8.03E−05</td>
<td>&lt;0.1%</td>
<td>75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver Female</td>
<td>9.67E−03</td>
<td>1.2%</td>
<td>72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liver Male</td>
<td>9.84E−07</td>
<td>&lt;0.1%</td>
<td>77%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muscle All</td>
<td>0.01</td>
<td>2.4%</td>
<td>81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muscle Female</td>
<td>0.05</td>
<td>4.3%</td>
<td>81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muscle Male</td>
<td>5.47E−03</td>
<td>0.7%</td>
<td>80%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
involved in both local and distant gene regulation. Interestingly, we found such highly conserved non-coding sequences within the CNV, suggesting a potential role for CNVs and/or regulatory elements within CNVs in the context of eQTL hotspots.

DISCUSSION

In this study, we have employed a combined genomics and genetics approach to ask whether CNVs in the mouse genome have a significant impact on gene expression levels and metabolic phenotypes. To this end, we have used microarray gene expression as well as metabolic traits measured in an F2 population obtained from an intercross between mouse inbred strains C57BL/6J and C3H/HeJ. We have found that CNVs play a significant role in gene expression and clinical traits in mice. We observed that 83% genes (45/54) found in CNV regions between B6 and C3H were differentially expressed (Table 1). These genes include a number of genes known to play a role in disease susceptibility in mouse and humans (22–27). For example, Glo1 is associated with autism susceptibility in humans (24), Alzheimer’s disease (27) and anxiety (25) in mice. Furthermore, Cfh is associated with age-related macular degeneration in humans (23) and reduced visual perception (26) in Cfh−/− mice. Interestingly, C3H mice, which have both lower Cfh expression and copy number, become blind as they age despite the addition of the gene Pde6b traditionally believed to cause blindness in C3H mice (28).

We observed that gene expression was generally increased in genes found in higher copy CNVs and decreased in genes found in lower copy CNVs (Table 2 and Supplementary Material, Table S4). However, this was not always the case, since ~20% of genes were expressed in the opposite direction as the copy number change, and roughly 15% of genes found in CNVs were not differentially expressed. There are several reasons that could explain these observations. First, gene expression may be influenced by additional regulators or transcription factors controlling gene expression in trans. Similarly, we cannot exclude the possibility that gene expression may be disrupted if regulatory regions are affected by the CNV or the possibility of transcriptional silencing due to DNA methylation. Another possibility involves the reliability in CNV detection, particularly in defining CNV boundaries with confidence when aCGH data are used (2). Finally, our analysis suggests tissue specific expression. We observed a larger number of genes whose expression was concordant with copy number in tissues which are largely homogenous, such as adipose tissue, but a lower degree of concordance in tissues with highly specialized sections, such as the brain (Table 2 and Supplementary Material, Table S4).

Our results also suggest that CNVs can influence complex phenotypes. We observed that several genes present in CNVs were highly correlated to metabolic traits such as body weight and adiposity (Fig. 3 and Supplementary Material, Table S6) and that cQTL for these traits map to CNVs. In addition, Klra8 was highly correlated with coronary artery calcification and Csf2ra was correlated with body...
weight, fat mass and insulin levels (Supplementary Material, Table S6). Both Klra8 and Cs2ra have been shown to play a role in immune-related functions such as in cytomegalovirus resistance (29) and immune cell differentiation, respectively. Interestingly, we find evidence that CD244 was correlated with femoral bone mineral density ($r = -0.38$, Supplementary Material, Table S6), a gene recently linked to rheumatoid arthritis (30), an autoimmune condition associated with bone loss in arthritis patients (31).

cQTL mapping to CNVs suggests that CNV genes, or regulatory elements within the CNVs, are contributing to trait development. The CNV genes Itlna, Trim12 and Trim34 were correlated to weight, triglycerides, adiposity, glucose and insulin levels (Fig. 3 and Supplementary Material, Table S6), and several metabolic traits show peak QTL mapping at or very near the location of these genes (Fig. 3). In addition, our causality test also suggests that transcript levels in the genes Trim12 and Trim34 are causal for plasma cholesterol and that Itlna and Gvin1 transcript levels are causal for plasma insulin levels, among others (Table 3). However, although this genetic and statistical evidence allows us to generate strong hypotheses for the involvement of CNV genes in clinical traits, the use of mouse knock-outs and transgens is still necessary to validate these causal relationships.

Recent studies have shown that both inherited and de novo CNVs are associated with autism (11,32,33) and schizophrenia (34,35) in humans, suggesting that CNVs may play a role in the etiology of complex traits. However, a deeper understanding of complex traits in humans necessitates the use of model organisms that can permit investigation of the molecular mechanisms underlying these traits. The vast number genetic tools available to the scientific community, as well as the availability of genome-wide gene expression profiles and extensive behavioral phenotyping databases (36,37), make the mouse an ideal choice for the dissection of CNVs. But can the mouse be used as a model to study CNVs? Our results suggest that it can. Even in the relatively small set of 19 CNVs variable between strains B6 and C3H, we observed a significant impact on gene expression levels in cis (Table 1) and possibly a much larger effect on genes regulated in trans (Fig. 1A, Fig. 5 and Supplementary Material, Fig. S2). Furthermore, the presence of highly conserved non-coding sequences in mouse CNVs (Fig. 5) suggests that these CNVs carry regulatory elements that influence the expression levels of tens or hundreds of eQTL in trans. We believe that our current work can serve as a starting point for the use of the mouse as a model to dissect the contribution of CNVs in complex traits.

### MATERIALS AND METHODS

**Copy number variation dataset**

A set of CNVs in 20 mouse inbred strains were previously identified by Graubert et al. (5) using aCGH. A set of 19 CNVs were identified between strains B6 and C3H. The putative genomic start and end positions of each CNV were obtained from the published data (5) and aCGH intensity data were obtained from the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) accession number GSE5805.

**Mouse crosses**

We analyzed two independent mouse intercrosses between strains C57BL/6J (B6) and C3H/HeJ (C3H) previously generated in the Lusis laboratory by crossing the parental strains to generate F1 mice, and then further intercrossing F1’s to generate F2 mice. The first cross was performed between strains B6 and C3H on an ApoE$^{-/-}$ background. Three hundred and forty-four F2 mice, 166 females and 168 males, were generated and fed a chow diet (Ralston-Purina Co, St Louis, MO, USA) until 8 weeks of age, then fed a high fat ‘western’ diet (Teklad 88137, Harlan Teklad, Madison WI, USA) for 16 weeks until euthanasia at 24 weeks of age. The second cross was performed between the same strains B6 and C3H on a wild-type background to generate 309 F2 mice, 145 females and 164 males. Mice were fed a chow diet until 8 weeks of age, and then fed a high fat ‘western diet’ for 12 weeks until euthanasia at 20 weeks of age. Since the second cross consisted of both B6×C3H and C3H×B6 F2 mice, we restricted our analysis to the B6×C3H mice, consist-
ent with the direction of the first cross. A detailed description of each cross is found in the articles published by Wang et al. (17) and Farber et al. (20). We referred to these crosses as cross 1 and cross 2, respectively. All mice were housed under specific pathogen-free conditions and according to NIH guidelines.

### Gene expression

RNA expression levels were measured by microarray using total RNA. Microarray analysis was carried out in adipose, brain, liver and muscle tissues from parental strains or F2 mice as described (17,20). Gene expression data are available on GEO for adipose (GSE3086), brain (GSE3087), liver (GSE2814) and muscle (GSE3088) tissues in the first cross, and in the second cross for adipose (GSE11065), brain (GSE3087), liver (GSE12795) and muscle (GSE12798) tissues.

### Cis–trans test

We used adipose RNA from B6, C3H and B6×C3H F1 mice from cross 1 to generate cDNA (ABI 4367381). We used PCR amplification on the cDNA using the following primers for the genes Krlk1 (5′ggg ata ggt act atc caa gta caa ggt3′ and 5′ctc ccc cat cct cag tta tct ct3′) and CNV14 (5′gca ggt gct cgg ctg atc gca aat ca3′ and 5′agg agg cat gta cgt agc cat cc3′). We used the single copy gene β-Actin to normalize each sample (5′gac cat gta cgt agc cat cc3′ and 5′ctc tca gct gtg gtt gta aa3′).

### QTL mapping

Expression and clinical QTL mapping was performed using the Rqtl package in R (www.rqtl.org). We used marker regression without imputation, and 95% confidence intervals were determined using a 1.5 LOD drop.

### NEO causality test

We used the NEO package in R to test causality using the simple model M → A → B, where A is clinical traits and M is the genotype of the nearest SNP to the gene. We tested each CNV gene–trait pair using expression and clinical trait data from cross 1, and selected causal gene–trait pairs where LEO scores >1 and RMSEA scores <0.05, as described (19,20).

### Conservation

We used the VISTA genome browser (http://pipeline.lbl.gov/) to determine the degree of conservation between mouse and human sequences using the genomic location of mouse CNVs. We used mouse Build 34 as the reference genome and determined conservation to human sequences as described (39).

### Statistical analysis

**Gene expression.** For each gene, we used SNP genotyping to determine the parental origin of a chosen genomic segment in the F2 mice, which allowed us to separate mice into three distinct groups, those carrying genomic segments homozygous for B6, homozygous for C3H or heterozygous. To determine if genes were differentially expressed, we used one-way ANOVA to compare microarray gene expression levels between B6 homozygous, C3H homozygous and heterozygous.
groups. To determine whether B6 homozygotes or C3H homozygotes were higher in gene expression levels, one-tailed right and left-handed \( t \)-tests were used to compare the means between the two groups.

**Binomial test.** To test the effect of CNVs on gene expression levels, we used the binomial probability of observing both a CNV and a change in gene expression levels for genes overlapping CNVs, given the genome-wide likelihood of observing gene expression differences. Three inputs were used for the binomial test: (i) the probability of a given gene being differentially expressed, \( P = \frac{\text{number of genes where the ANOVA test gave } P\text{-value less than 0.05}}{\text{total number of genes in microarray}} \), (ii) the total number of genes overlapping CNV regions, \( n \), and (iii) the total number of ‘successes’, genes which overlap CNVs and were also differentially expressed, \( x \). We took as \( P\)-value 1 minus the binomial cumulative distribution function, with parameters \( p \), \( n \) and \( x \).

We also tested whether copy number was concordant with gene expression; for example, so that if a CNV was higher in B6, the level of gene expression was also higher in B6. We again used the binomial test, where number of successes \( x \) was defined as the number of genes where the level of expression was both higher in B6 relative to C3H and the CNV change was also higher in B6, plus the number of genes where the level of expression was both higher in C3H and the CNV change was also higher in C3H. Each binomial test was performed using the `binocdf` function in MATLAB.

**Gamma test.** For CNVs that overlapped more than one gene, we tested the effect of the CNV on gene expression as follows: if a given CNV overlaps \( x \) genes, and the probability that each \( i \)th gene is differentially expressed is \( p_i \), then in the random model that the \( p \) are uniform on \((0,1)\), \( a = p_1 p_2 \cdots p_x \) has a cumulative distribution function \( \text{gamma}(\log(a), x, \text{‘upper’}) \), in MATLAB notation. A regularized incomplete gamma function (40) serves as a suitable \( P\)-value for the probability that at least one of the genes is differentially expressed.

**Correlation.** We correlated gene expression levels of genes overlapping CNVs to metabolic quantitative traits measured in each F2 mouse. Metabolic traits were measured as described (17,20). For each gene–trait pairing, Pearson’s correlation coefficient \( r \) was calculated for the vector of gene expression values and corresponding vector of trait values across F2 mice. For each gene–trait correlation, we computed \( r \) using only F2 mice were both gene expression and trait data were available for the given gene–trait pair.

**Overlap between replication crosses.** We used a hypergeometric test to determine the overlap in the set of genes hypoth-
Figure 5. Mouse eQTL hotspot coincides with highly conserved non-coding sequences present in mouse CNVs. (A) A CNV in chromosome 3 coincides with the 95% confidence interval of 27 eQTL. Each curve represents an eQTL with LOD score >4.3 from female liver in cross 1. The red bar marks the position of the CNV. (B) The location of the CNV and neighboring genes in a 2 Mb window from the UCSC genome browser in Build37. Cross-species conservation for the CNV sequences is shown in (C), where the base mouse genome sequence (Build 37) is compared with the rat, human, rhesus macaque, dog, horse and chicken sequences. The degree of conservation is plotted on the y-axis as % identity over a 100 bp window. Highly conserved regions are highlighted under the curve, with a red color specifying non-coding sequences.
sized to be influenced by CNVs in the two mouse crosses. We employed the hypergeometric test to obtain P-values for the intersection of the sets (i) the CNV genes differentially expressed by ANOVA in cross 1 and (ii) the CNV genes differentially expressed by ANOVA in cross 2, being unusually large in the universe of all CNV genes.

Furthermore, we tested whether the directionality of gene expression differences was consistent between the two crosses using Pearson’s correlation. The overlap in the genes identified in each cross was calculated in adipose, brain, liver and muscle tissues separately.

Multiple comparisons. To determine false discovery (e.g. due to multiple testing), we permuted mouse and genotype relationships and repeated ANOVA, t-test, binomial test and gamma test panels 1000 times. We estimated the rate of false discovery by comparing the P-values obtained from the unpermuted (‘true’) data to their relative rank in the distribution of P-values from the permuted datasets. If the P-value was smaller than all permuted P-values, we assigned a false discovery of <0.1% due to the limited resolution inherent in using 1000 permutations. To assign a false discovery rate for the gene–trait correlations, we used the modified Benjamini and Hochberg FDR approach described by Storey (18) and selected a 1% FDR cutoff based on the correlation P-value distributions in each tissue.

All statistical analyses were performed using MATLAB.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG online.

ACKNOWLEDGEMENTS

We thank Eric Schadt for his help in eQTL mapping, Rosetta Inpharmatics for funding of expression microarrays and Hannah Qi, Xuping Wang and Judy Wu for their help in tissue collection and trait measurements.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the USPHS National Research Service Award GM07104; National Institutes of Health training grant 5T32HD07228 and program project grant NIH/NHLBI HL28481 and HL30568.

REFERENCES