Supplementary Methods:

Fine mapping of regulatory loci for mammalian gene expression using radiation hybrids

Christopher C. Park[1], Sangtae Ahn[2], Joshua S. Bloom[1,7], Andy Lin[1], Richard T. Wang[1], Tongtong Wu[3,7], Aswin Sekar[1], Arshad H. Khan[1], Christine J. Farr[4], Aldons J. Lusis[5], Richard M. Leahy[2], Kenneth Lange[6] and Desmond J. Smith[1,8]

[1] Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

[2] Department of Electrical Engineering, Signal and Image Processing Institute, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA

[3] UCLA School of Public Health, Department of Biostatistics, Los Angeles, CA 90095, USA

[4] Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK

[5] Department of Microbiology, Immunology and Molecular Genetics, Department of Medicine, and Department of Human Genetics, University of California, Los Angeles, CA 90095, USA

[6] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA

[7] Current addresses: Department of Molecular Biology, Princeton University, Princeton, NJ 08544 (JSB), Department of Epidemiology and Biostatistics, College of Health and Human Performance, University of Maryland, College Park, MD 20742 (TW)

[8] To whom correspondence should be addressed. Email: DSmith@mednet.ucla.edu

Comparative genomic hybridization genotyping

For each hybrid clone, we harvested two batches of cells separated by a freeze/expansion cycle. The hybrids were re-genotyped using DNA from the second batch of cells and mouse comparative genomic hybridization (CGH) arrays from Agilent. DNA from the A23 recipient hamster cell line served as the control channel. The CGH arrays displayed 60-mer oligonucleotides in situ synthesized by ink-jet printer technology[1]. The length of the oligonucleotides helped ensure that DNA from both the donor mouse and endogenous hamster genomes were detected (see also Transcript analysis, below). DNA was labeled using the Agilent Genomic DNA Labeling Kit Plus; Cy5 was used for the RH cell lines and Cy3 for the A23 cell line. Labeled DNA was applied to the Agilent G4415A Mouse Genome 244k CGH microarray and scanned according to manufacturer's instructions. The translation between the Agilent ID reference and genome co-ordinates can be found on the Agilent web site.

Transcript analysis

Total RNA was extracted from the RH cell lines using the Qiagen Mini RNeasy kit. Two biological replicates of RNA were prepared, from the cells grown before and after the freeze/expansion cycle (Comparative genomic hybridization genotyping, above). RNA was also prepared from the recipient hamster cell line A23. The RNA was converted to labeled cDNA using the Agilent Fluorescent Direct Label Kit as per manufacturer's instructions. Labeled products were applied to the Agilent G4121A mouse oligo microarray with a dye swap. For the first replicate, the RH RNA was labeled with Cy5 and the A23 RNA with Cy3. For the second replicate, the labeling was reversed. The

translation between the Agilent ID reference and gene names can be found on the Agilent web site.

The Agilent microarrays displayed 60-mer oligonucleotides. These long oligonucleotides allow medium abundance transcripts with 6 nucleotide mismatches to bind with > 90% efficiency[1]. Our own analyses confirmed that RNA from both the donor mouse and endogenous hamster genes were detected with comparable efficiency (Mouse/hamster sequence conservation, below). Thus, regulation of either recipient hamster or donor mouse genes by extra donor gene copies could be evaluated.

Preprocessing CGH data

The CGH data (RH/A23 reference) for the 232,626 markers was $\log_{10}$ transformed (see Preprocessing expression data, below). The distribution was bimodal (**Supplementary Fig. 1b**). The first mode represented regions with no extra copy of a mouse gene and the second mode regions with one or more extra copies. The proportion of clones retaining two copies of the donor genome is small (~the square of the retention frequency or $(0.239)^2$, i.e. 5.7%). Because of this low frequency, clones with two donor copies were not clearly resolved as either a shoulder or third mode in the histogram.

The CGH data was analyzed as $\log_{10}$(RH/A23) copy number ratio averaged over a sliding window of 10 markers. This window size gave the best compromise between reducing data variance while producing least degradation in resolution.

The CGH data were normalized as follows. First, the data for each array were subtracted by its own individual mode (a CGH value corresponding to the first peak). This small correction gave the first mode for each array a $\log_{10}$(RH/A23) copy number ratio of zero. In the next step, the CGH data was pooled over all arrays and the first and second modes estimated by modeling the data as a mixture of two Gaussian distributions (see also Mixture models for CGH data, below)[2,3]. Because extra copy regions in the autosomes represent three copies (one mouse and two hamster copies) compared to two (two hamster copies), the data for the autosomes was multiplied by a common scaling factor of ($\log_{10}$[3/2])/(secondary mode CGH value - primary mode CGH value). Consequently, the normalized CGH data for the autosomes had a primary mode at $\log_{10}$(2 copies/2 copies) = 0 and a secondary mode at $\log_{10}$(3 copies/2 copies) = 0.176. Since the A23 hamster recipient cells are male, extra copy regions on the X chromosome represent two copies (one mouse plus one hamster) compared to one (one hamster copy). The data for the X chromosome was therefore multiplied by $\log_{10}$(2/1)/(secondary mode CGH value - primary mode CGH value), giving a secondary mode at $\log_{10}$(2 copies/1 copy) = 0.301.

Concordance between PCR and CGH markers

We evaluated the concordance between PCR and CGH markers in the hybrids by averaging the $\log_{10}$(RH/A23) copy number ratio of the five neighboring CCH markers to the left and right of each PCR marker. If the average signal of the ten CGH markers exceeded 95% of the distribution representing the first (baseline) mode, the local region was classified as being retained in a particular clone.

Retention frequency

To assess the retention properties of the RH panel, an individual marker was classified as being retained in particular clone if its CGH signal exceed 95% of the baseline distribution. Similar results were obtained using a mixture model (Mixture models for CGH data, below).

Preprocessing expression data

The expression data was normalized using GeneSpring (Agilent Technologies). Each gene value was divided by the control channel and each chip normalized to the 50^{th} percentile of the measurements taken from that chip. The $\log_{10}$(RH/A23) expression ratios for the 20,145 genes were averaged over the two dye-swapped arrays. The log transformation was employed because the non-log transformed data had a large dynamic range (10^{-2}-$10^{4.3}$) and a large standard deviation (~33) compared to its mean (~1.5)[4]. The transformation dramatically improved the normality of the data and is how microarray expression ratios are usually treated (for example, see refs. 5,6). All outliers with expression > 5 standard deviations from the median were removed from the data after normalization.

Mouse/hamster sequence conservation

To evaluate the relative efficiency of the arrays in detecting mouse and hamster transcripts, we co-hybridized mouse and hamster brain RNA to two separate arrays using a dye swap. Brain was chosen because most genes are expressed in this tissue.

Additional genes were surveyed by repeating the experiment using liver RNA on another two arrays, also with a dye swap. As anticipated, there was a high correlation between the mouse and hamster expression signals, suggesting strong sequence conservation between the two species (brain, R = 0.867, $P < 2.2 \times 10^{-16}$; liver, R = 0.868, $P < 2.2 \times 10^{-16}$; averaged brain and liver, R = 0.863, $P < 2.2 \times 10^{-16}$) (**Supplementary Figs. 2a-c**). There was a modest excess of points above the regression lines, consistent with more efficient detection of mouse than hamster sequences for those genes.

The normalized $\log_2$(mouse/hamster) expression ratio averaged across brain and liver (four arrays total) was used to evaluate sequence differences between the two species. As expected from the correlation coefficients, the distribution of the averaged $\log_2$(mouse/hamster) expression ratios was unimodal with a median close to zero (0.04 $\pm$ 0.008) and a modestly extended right tail (**Supplementary Fig. 2d**).

In addition to the expected modest excess of mouse transcripts binding more strongly to the array than hamster, there was also a small minority of genes (~7 % with average $\log_2$(mouse/hamster) expression < -1) where the hamster transcripts appear to bind better than mouse. Possible explanations are (1) the inevitable noise in microarray measurements (2) some genes may be more highly expressed in hamster liver or brain than mouse and (3) sequence errors which happen to match hamster better than mouse. In any case, differential binding of mouse and hamster transcripts cannot mimic trans regulation.

We also examined the limited amount of available hamster sequence to identify conservation with the oligonucleotides on the Agilent expression array. In the available 152 overlaps, the mouse/hamster sequences were 89% conserved (**Supplementary Fig. 3a**). As expected, there was a significant correlation between the mouse/hamster $\log_2$ expression ratio for each of the 152 probes and the number of mismatches between the two species (R = 0.25, P = 0.002) (**Supplementary Fig. 3b**). Based on the limited data, there was no obvious position in the 60-mer where a mismatch gave a dominant effect on hybridization (**Supplementary Fig. 3c**). Overall, the data indicated that the arrays can detect both mouse and hamster transcripts with comparable efficiency.

Models

The T31 RH panel can be usefully viewed as a library of cells containing many random donor fragments, allowing the effects of extra gene copies to be evaluated for the whole genome. This approach provides greater efficiency than employing a cell line for each gene and allows the statistical association between the effect of every marker on gene expression to be evaluated. The independent retention of multiple fragments also allows combinatorial models of interacting genes to be explored.

We used two regression models to analyze the data (**Fig. 3a**). For both models the mean value was conveyed by the parameter μ, which takes the potential effect of other ceQTLs into account. Significance was indicated by $-\log_{10}P$.

Model 1 regressed the expression of each gene on each marker, thus identifying trans ceQTLs for distant markers (> 10 Mb from a gene) and cis ceQTLs for local markers (< 10 Mb)[7]. The effect size was conveyed by the parameter α. An F-test was used to evaluate each marker and gene pair for statistical association[4]. For a marker and gene, we fitted a full model

$$t = \mu + \alpha x$$

over all RH clones where t is the normalized $\log_{10}$(RH/A23) expression data and x is the normalized $\log_{10}$(RH/A23) CGH data. The estimated regression coefficients from the least squares method are given by

$$\hat{\alpha} = \frac{\sum (t_i - \bar{t})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\mu} = \bar{t} - \hat{\alpha}\bar{x}$$

where $\bar{t} = \sum t_i / n$ and $\bar{x} = \sum x_i / n$, with $n = 99$ being the number of RH cells. To test the null hypothesis that $\alpha = 0$, we also fitted a reduced model

$$t = \mu$$

where the least squares estimate of μ is simply given by

$$\hat{\mu}^* = \bar{t}\ .$$

The standard F-statistic is given by

$$F = \frac{SSE(RM) - SSE(FM)}{SSE(FM)/(n-2)}$$

where the sum of squared errors (SSE) for the full model (FM) is given by $SSE(FM) = \sum (t_i - \hat{\mu} - \hat{\alpha} x_i)^2$ and SSE for the reduced model (RM) $SSE(RM) = \sum (t_i - \hat{\mu}^*)^2$.

Model 2 evaluated the interaction between local and distant markers on the expression of a gene. This model hence explored whether a distant marker affected the expression of two hamster copies of a gene differently from two hamster copies plus an extra mouse copy. Significant results would yield ceQTL peaks in which the -$\log_{10}P$ value referred to the interaction between local and distant markers. Note that Model 2 can identify interacting loci distinct from trans ceQTLs in Model 1. The effect size for Model 2 was conveyed by the parameter γ. The CGH signal for local markers in Model 2 was obtained by linear interpolation of the pair of closest markers to the 5' and 3' of each gene. To test distant-local marker pairs for interaction, we also performed an F-test. We first fitted a full model

$$t = \mu + \alpha x + \beta y + \gamma xy$$

where x and y represents the $\log_{10}$ transformed CGH data for each of the distant and local marker pairs, respectively. We then computed $SSE(FM) = \sum (t_i - \hat{\mu} - \hat{\alpha} x_i - \hat{\beta} y_i - \hat{\gamma} x_i y_i)^2$ where $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are least squares estimates of the regression coefficients[4]. To test the null hypothesis that $\gamma = 0$, we also fitted a purely additive reduced model

$$t = \mu + \alpha x + \beta y$$

and computed $SSE(RM) = \sum (t_i - \hat{\mu}^* - \hat{\alpha}^* x_i - \hat{\beta}^* y_i)^2$ where $\hat{\mu}^*$, $\hat{\alpha}^*$ and $\hat{\beta}^*$ are least squares estimates of the regression coefficients. The F-statistic was given by

$$F = \frac{SSE(RM) - SSE(FM)}{SSE(FM)/(n-4)}.$$

In addition to Model 1, we also used a weighted regression model to evaluate the expression data (Weighted regression model, below)[8]. To implement this model, posterior probabilities for whether a marker represented 1 or 0 mouse copies based on the CGH data were computed using a mixture model (Mixture models for CGH data, below)[2]. We found very good agreement between the effect size α for the weighted regression model and Model 1. Because of its computational simplicity, we employed Model 1 for the analyses in the paper.

Calculating *P* values and FDRs for Models 1 and 2

To calculate the *P* value $P(F)$ for an observed statistic F in Model 1, the null distribution of the test statistic is needed. Instead of assuming that the test statistic followed the F-distribution, we estimated the null distribution by permutation[9]. The expression data were randomly permuted five times and F-statistics recalculated (giving 5 x 20,145 x 232,626 samples), as described[6,10]. The recalculated F-statistics were pooled to obtain an empirical null distribution and *P* values calculated as the frequency of null statistics exceeding the observed statistic.

Note that for each gene-marker pair, this scheme permutes the expression data for the gene and is equivalent to permuting the genotyping data for the respective marker. The strategy thus maintains the correlation structure between gene expression traits and accounts for spurious trans ceQTL hotspots. The permutations were run with and without the two regions of extreme retention frequency at the *p53* gene (<5% retention) and the *tk* gene (>95% retention) both on chromosome 11. There was no detectable difference between the resulting empirical significance thresholds, reflecting the small number of extreme markers.

To control false discovery rates (FDRs), FDR-adjusted *P* values (*Q* values) were computed following the Benjamini-Hochberg procedure[11,12]. If a marker has a significant effect on the expression level of a gene, the marker is part of a copy number expression quantitative trait locus (ceQTL) for the gene. A marker is part of a trans ceQTL if the distance between the marker and the regulated gene is > 10 Mb. The ceQTL is in cis, otherwise. In Model 1 the FDR procedure was performed separately for the cis and trans ceQTLs, since cis ceQTLs do not require genome-wide significance thresholds[13].

For Model 2, the distribution of the null F-statistics was determined using the residual empirical method and permutation[14]. Five permutations of the expression data were carried out and the results pooled to obtain a null distribution[6]. The *P* value was calculated as the frequency of null statistics exceeding the observed statistic. The FDR-adjusted *P* values (*Q* values) were calculated by the Benjamini-Hochberg procedure.

Mixture models for CGH data

Preparatory to a weighted regression model of the expression data (Weighted regression model, below), we used a finite mixture model[2,3] to evaluate each RH clone for the probability of an extra copy of a mouse gene based on the CGH data. For each array a mixture of two Gaussian distributions was fitted to the data as follows:

$$f(y) = p_0\phi(y;u_0,\sigma_0^2) + p_1\phi(y;u_1,\sigma_1^2)$$

where y is the $\log_{10}$(RH/A23) CGH data, f is the probability density function (pdf) for y, $\phi(\cdot;u,\sigma^2)$ denotes the pdf of the Gaussian distribution with mean u and variance σ^2, p_0 and p_1 are the mixture weights, u_0 and σ_1^2 are the parameters for the first mode representing two copies (two hamster copies) and u_1 and σ_1^2 are the parameters for the second mode representing three copies (two hamster copies and one mouse copy), such that $u_0 < u_1$. An expectation-maximization (EM) algorithm[2,3] was used to compute the maximum likelihood estimates of the parameters p_0, u_0, σ_0^2, p_1, u_1 and σ_1^2. Using the estimated parameters, the probability of the presence of an extra copy, that is, three copies, was calculated by

$$\tau(y) = \frac{p_1\phi(y;u_1,\sigma_1^2)}{p_0\phi(y;u_0,\sigma_0^2) + p_1\phi(y;u_1,\sigma_1^2)}.$$

The probability of two copies follows as $1-\tau$. In actual computation τ was enforced to be monotone in y, implying that the larger the CGH signal, the more likely the presence of an extra copy. When $\tau > 0.5$, a CGH marker in an RH clone was called as having an extra copy. The retention frequency obtained using the mixture model for the T31 RH panel was 20.2 $\pm$ 0.02%, similar to the estimate in the paper, 23.9 $\pm$ 0.02% (see also

Retention frequency, above). In addition, there was a high level of agreement between the two estimated retention frequencies for all markers (R = 0.96, $P < 2.2 \times 10^{-16}$).

Weighted regression model

A weighted regression model[8] was fitted to the expression data for each gene as follows:

$$t_i = \mu + \alpha z_i + \varepsilon_i$$

where t_i is the normalized $\log_{10}$(RH/A23) expression data in RH clone i, z_i is a Bernoulli random variable with success probability $\Pr(z_i = 1) = \tau_i$, ε_i is the residual error with $N(0,\sigma^2)$, and μ and α are unknown parameters representing a baseline level and an effect size, respectively. That $z_i = 1$ means the presence of an extra copy, and the success probability τ_i was calculated from the mixture model (Mixture models for CGH data, above). An iteratively reweighted least squares (IRWLS) method[8] was used to fit the weighted regression model and to estimate the parameters. The estimated α parameters from the simple regression model (Model 1) and the weighted regression model were highly correlated (R = 0.8934, $P < 10^{-300}$) (**Supplementary Fig. 4a**).

Replicability

To assess the quality of the data, we examined the two replicate datasets individually using Model 1. Overall, the correlation coefficient of the α values between the two replicate datasets was 0.887 ($P < 2.2 \times 10^{-16}$) for all FDRs < 0.4 (**Supplementary Fig. 4b**). At the same FDR of < 0.4, there were 6,331,188 markers above threshold from the first dataset and 8,127,330 from the second. Although the overlap of 2,963,068

(36.5%) was limited, reflecting the noisy nature of array data, it was nevertheless strongly significant (χ^2 = 7.96 x 10^8, df = 1, $P < 10^{-300}$). The degree of overlap grew with decreasing FDR (**Supplementary Table 1**). These observations were obtained using biological replicates of the RH clones after a freeze/expansion cycle and indicate the data is of good reproducibility. A trans ceQTL derived separately from the two datasets is shown in **Figs. 5e** and **5f**.

Cis ceQTLs

By co-hybridizing mouse and hamster RNA, we had shown that the expression arrays detected transcripts from both species with comparable efficiency (Mouse/hamster sequence conservation, above and **Supplementary Figs. 2** and **3**). Nevertheless, some of the cis ceQTLs in Model 1 might stem from sequence differences between the two species. In this situation, the mouse gene would give a stronger signal than its hamster ortholog when present in the RH clones. In classical eQTL mapping, nearly half of cis eQTLs may be artifacts due to sequence polymorphisms affecting hybridization efficiency[15]. We therefore compared the distribution of the average $\log_2$(mouse/hamster) expression ratios (**Supplementary Fig. 2d**) for genes regulated by cis ceQTLs and trans ceQTLs. The trans ceQTL distribution is expected to be unaffected by mouse/hamster sequence differences and acts as a control. There was little difference between the two distributions, suggesting that the majority of cis ceQTLs are unrelated to sequence differences between the two species (**Supplementary Fig. 5**).

Replicability of negative α cis ceQTLs

Although the degree of overlap between the two replicate datasets was much poorer for the negative α cis ceQTLs at FDR < 0.4 (5.2% of markers) than the positive (48.4%), both overlaps were still significantly greater than chance (χ^2 = 5.84 x 10^4 and 7.88 x 10^6 respectively, df = 1, P < 2.2 x 10^{-16} each comparison). In addition, the correlation coefficient of the α values between the two replicate datasets was similar for the negative and positive α cis ceQTLs (R = 0.72 and 0.87 respectively, P < 2.2 x 10^{-16} for both). These observations suggest that the negative α cis ceQTLs may be driven partly by noise but at least some are replicable and not due to outliers (**Supplementary Figs. 6a-c**).

Hotspot analysis

We evaluated whether the number of genes regulated in trans by each ceQTL with FDR < 0.4 was significantly high[5]. The total sum of regulated gene/marker pairs for 232,626 markers was 1,162,130. If the 1,162,130 gene/marker pairs were randomly distributed across the 232,626 markers, the number of regulated genes for each marker would follow a Poisson distribution with a mean of 5.00. Using this null distribution, we calculated a one-sided P value for each observed number of regulated genes. We computed a corresponding Q value (FDR-adjusted P value) by the Benjamini-Hochberg procedure.

Highly regulated genes

We tested whether each gene was regulated by a significantly high number of trans ceQTLs with FDR < 0.4 following the procedure described for the hotspots (Hotspot analysis, above). The most highly regulated gene with FDR < 0.4 was aspartyl aminopeptidase (*Dnpep*) regulated by 42 trans ceQTLs. It is unlikely that expression variation observed in 99 RH clones can be meaningfully decomposed into as many as 42 explanatory ceQTLs, suggesting that simple regression may not be sufficiently conservative in this context.

One possible explanation for genes with large number of trans ceQTLs could be that non-normality and outliers in the expression of these genes might be diluted by the many other genes in the statistical treatment. To ensure that such effects were properly accounted for, we validated the global permutation scheme by implementing an individual gene-marker pair based permutation scheme for *Dnpep*, the most highly regulated gene. For each of the 232,262 markers, the expression data of the gene were randomly permuted 10,000 times, F-statistics recalculated (giving 10,000 samples) and *P* values calculated for each marker. The *P* values obtained from the global and the individual gene-marker pair permutation schemes agreed very well ($R = 0.9999$, $P < 10^{-300}$). Thus even though highly regulated genes may have a number of trans ceQTLs approaching the number of hybrids, such genes are still likely to be more highly regulated than average.

We examined the five most highly regulated genes (aspartyl aminopeptidase, *Dnpep*, regulated by 42 trans ceQTLs; vestigial like 3, *Vgll3*, 39 trans ceQTLs; myristoylated alanine rich protein kinase C substrate-like 1, *Marcksl1*, 38 trans ceQTLs; proline rich 13, *Prr13*, 37 trans ceQTLs and split hand/foot malformation (ectrodactyly) type 1, *Shfdg1*, 34 trans ceQTLs) for unusual properties in the SymAtlas database[16]. This publicly available database provides microarray expression data for all genes across 61 mouse tissues. The mean $\log_{10}$ expression value across mouse tissues for each of the five most highly regulated genes in the RH panel showed no obvious exceptional properties with values of 2.26 $\pm$ 0.01 (*Dnpep*), 1.73 $\pm$ 0.01 (*Vgll3*), 3.09 $\pm$ 0.08 (*Marcksl1*), 3.21 $\pm$ 0.05 (*Prr13*) and 4.13 $\pm$ 0.03 (*Shfdg1*) compared to 2.18 $\pm$ 0.003 (all genes). Similar results were found for the standard deviation across tissues, with values of 0.095 (*Dnpep*), 0.080 (*Vgll3*), 0.616 (*Marcksl1*), 0.381 (*Prr13*) and 0.214 (*Shfdg1*) compared to 0.181 (all genes).

Comparing RH and SymAtlas data

We compared the RH and SymAtlas data for the 15,220 genes overlapping in the two datasets. For each dataset, we constructed a correlation matrix whose element is the Pearson's correlation coefficient between the expression data for each pair of genes. We then tested whether the correlation matrices for the RH and the SymAtlas data were significantly more similar than expected by chance. For a dissimilarity measure, we used the Frobenius norm of the difference of the two matrices. The Frobenius norm $\|A\|_F$ of matrix A is given by

$$\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$$

where a_{ij} are elements of A. The distance (or dissimilarity) between matrix A and B is given by $dis(A, B) = \|A - B\|_F$. To test whether the distance between the correlation matrices was significantly small, we obtained the null distribution by calculating the distance between the RH correlation matrix and the randomly permuted SymAtlas correlation matrix. A total of 10,000 permutations were performed. The inferred *P* value was the frequency of permutations giving a difference less than the observed difference. The resulting *P* value was < 10^{-4}.

We also asked whether the genes regulated by the top hotspot showed higher correlations than chance in the SymAtlas dataset. Of the 614 genes regulated by the chromosome 5 hotspot, 486 were found in SymAtlas. A permutation t-test was performed comparing the mean pairwise correlation coefficients in SymAtlas of these genes with the null distribution obtained by sampling from all gene pairs in common with the RH dataset. The genes regulated by the hotspot showed a significantly greater correlation in mouse tissues than random (*P* = 1.6 x 10^{-3}), again suggesting that regulation of gene expression in the RH panel is similar to that occurring in the normal mouse.

Permutation t-tests

Permutation t-tests were used to evaluate the significance of differences between two groups in which the observed group was a subset of the parental group. The null distributions for the tests were obtained by taking the mean of random samples from the

parental group having the same number of observations as the tested group. The number of permutations was 5 x 10^4. The frequency of observations in the null distribution that exceeded the mean of the observed group was taken as the *P* value.

Transfection

We obtained a *Pcdh7* isoform a (*Pcdh7a*) cDNA (GenBank accession number BC131967), containing the entire coding region but lacking most of the 3' untranslated region (Open Biosystems). The full coding sequence was excised using the restriction enzymes MluI and EcoRV and inserted into the mammalian expression vector pCMVSport6 (Invitrogen) which provided the CMV promoter and SV40 3' untranslated sequences. The resulting construct was transfected into both HEK 293 cells and the A23 hamster recipient cells used for the RH panel. Cells were transfected using Effectene (Qiagen) lipofection reagent. Parallel transfections wee performed using the empty vector and a construct expressing GFP. Under fluorescence microscopy, greater than 40% of the cells were observed to be transfected with GFP in both the HEK and A23 cell lines. The cells were harvested after 48 hours and total RNA extracted. Two biological replicates were obtained for each sample (transfected and transfected with empty vector, HEK 293 and A23).

RNA from the transfection experiments was labeled using the Agilent Low RNA Input Linear Amplification Kit Plus and applied to the Agilent G4122A 44K mouse oligo microarray. The experimental RNA (from cells transfected with *Pcdh7a*) and control RNA (from cells transfected with an empty vector) were co-hybridized to the arrays and

scanned according to the manufacturer's instructions. The translation between the Agilent ID Reference and gene names can be found on the Agilent web site.

Real-time qPCR was used to assess overexpression of *Pcdh7a* because the construct lacked the 3' untranslated sequences recognized by the array. Compared to cells transfected with empty vector, the *Pcdh7a* isoform was overexpressed ~500-fold in transfected cells. Primers and probes were obtained from Applied Biosystems. Pdch7a overexpression in the RH panel (1.44-fold) was evaluated by comparing clones containing the gene to those lacking it, based on local CGH markers.

Trans ceQTLs lacking known genes

Known genes were defined as those appearing in the UCSC genome browser, Refseq, the Mammalian Gene Collection and the microRNA database miRBase. There were 4,485 unique trans ceQTLs of FDR < 0.4 and $-\log_{10}P > 4$ with no genes within a 150 kb radius from the peak marker, and 2,761 ceQTLs with no genes within a 300 kb radius (**Fig. 7a**). In this tally, a locus regulating one or more genes was counted as one trans ceQTL. To be conservative, we assigned trans ceQTLs as having no genes to those lacking genes within a 300 kb radius of the peak $-\log_{10}P$ marker. These trans ceQTLs are unlikely to be adversely affected by inflated discovery rates due to marker excess in non-gene regions, since only 13.3 % of randomly chosen markers lack a gene within a 300 kb radius.

The relation between the number of trans ceQTLs lacking genes and their $-\log_{10}P$ and FDR values are shown in **Supplementary Fig. 6d**. Examples of regression lines relating expression to peak marker CGH data for trans ceQTLs with both positive and negative α values lacking genes are shown in **Supplementary Figs. 6e** and **6f**.

URLs

Translation between Agilent G4415A mouse genome 244k CGH microarray ID reference and genome co-ordinates: http://www.chem.agilent.com/cag/bsp/oligoGL/014695_D_GeneList_20070207.txt.zip.

Translation between Agilent G4121A mouse oligo microarray ID reference and gene names: http://www.chem.agilent.com/cag/bsp/oligoGL/011978_D_GeneList_20050310.html.

SymAtlas database: http://symatlas.gnf.org/SymAtlas/. Open Biosystems: http://www.openbiosystems.com/. Applied Biosystems: http://www.appliedbiosystems.com/. UCSC genome browser: http://genome.ucsc.edu/. Refseq: http://www.ncbi.nlm.nih.gov/RefSeq/. Mammalian Gene Collection: http://mgc.nci.nih.gov/. miRBase: http://microrna.sanger.ac.uk/.

Mapping results

Mapping results together with FDRs are available upon request.

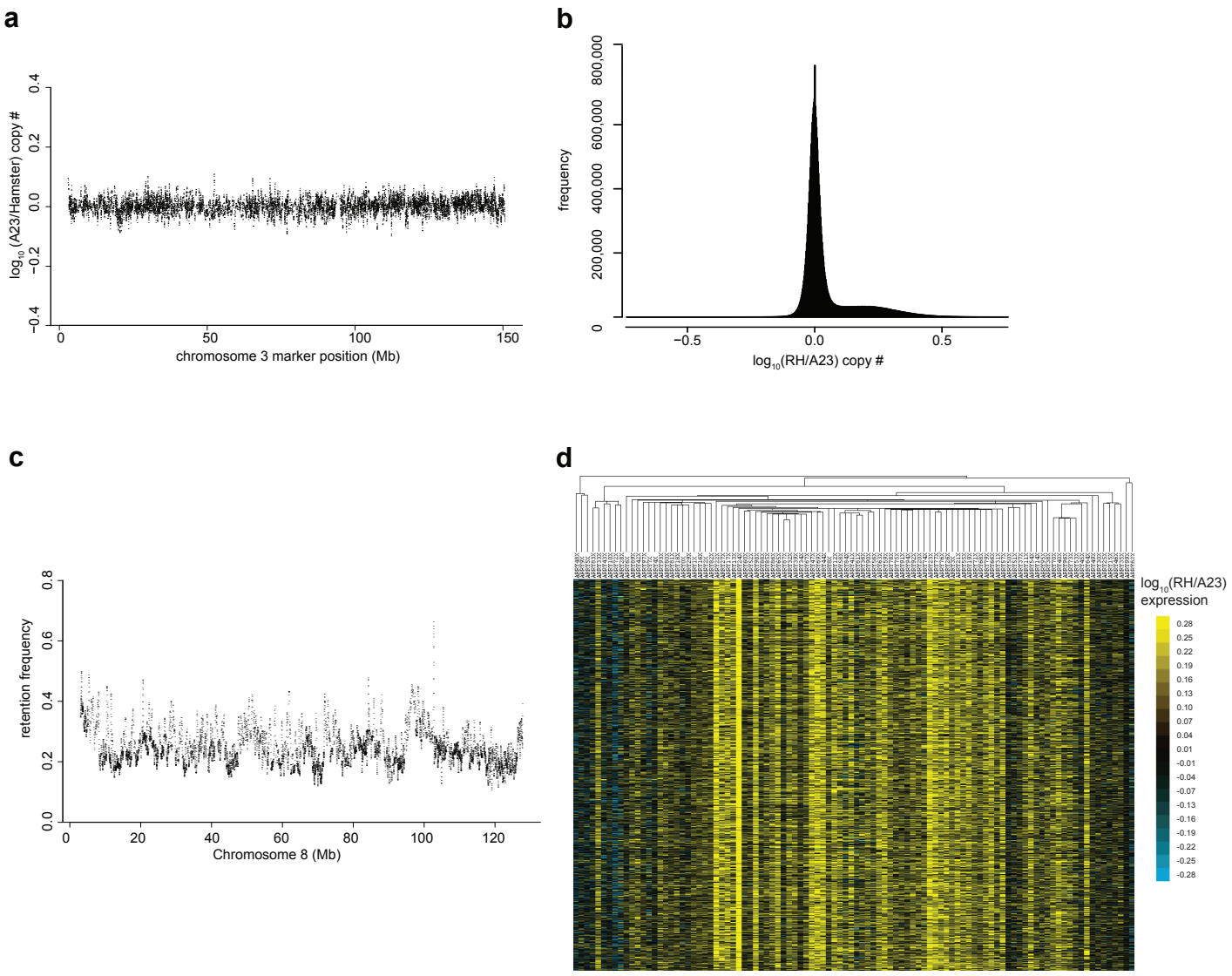
References

1. Hughes, T.R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* **19**, 342-347 (2001).
2. Jansen, R.C. Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* **49**, 227-231 (1993).
3. Redner, R.A. & Walker, H.F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* **26**, 195-239 (1984).
4. Chatterjee, S., Hadi, A.S. & Price, B. *Regression analysis by example*, (John Wiley & Sons, New York, 2000).
5. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* **102**, 1572-1577 (2005).
6. Storey, J.D., Akey, J.M. & Kruglyak, L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**, e267 (2005).
7. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**, 862-872 (2006).
8. Xu, S. Further investigation on the regression method of mapping quantitative trait loci. *Heredity* **80**, 364-373 (1998).
9. Churchill, G.A. & Doerge, R.W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971 (1994).
10. Brem, R.B., Storey, J.D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701-703 (2005).
11. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Royal Stat Soc, Series B* **57**, 289-300 (1995).
12. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat* **29**, 1165-1188 (2001).
13. Carlborg, O. et al. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics* **21**, 2383-2393 (2005).
14. Doerge, R.W. & Churchill, G.A. Permutation test for multiple loci affecting a quantitative character. *Genetics* **142**, 285-294 (1996).
15. Alberts, R. et al. Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE* **2**, e622 (2007).
16. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).

Supplementary Table 1. Overlap between biological replicate datasets

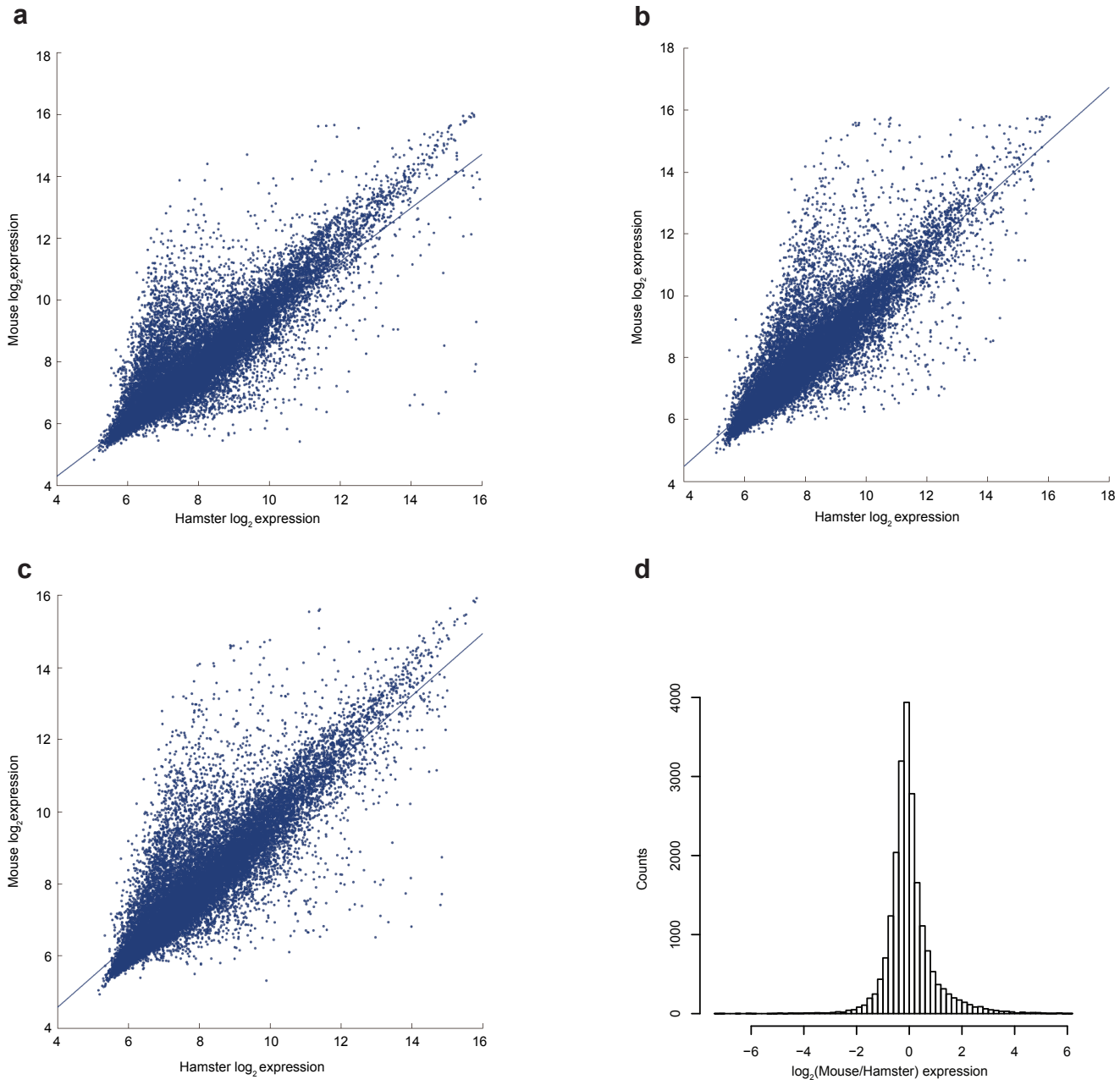
FDR	Significant markers dataset 1	Significant markers dataset 2	Overlap	% of dataset 2	Expected overlap	χ^2	df	*P* value
0.4	6,331,188	8,127,330	2,963,068	36.5	10,980	7.96×10^8	1	$< 10^{-300}$
0.2	3,366,266	4,253,375	1,906,959	44.8	3,055	1.19×10^9	1	$< 10^{-300}$
0.1	2,270,819	2,902,372	1,411,134	48.6	1,406	1.41×10^9	1	$< 10^{-300}$
0.05	1,672,038	2,164,346	1,093,842	50.5	772	1.55×10^9	1	$< 10^{-300}$
0.01	942,079	1,245,554	649,044	52.1	250	1.68×10^9	1	$< 10^{-300}$

Figure S1



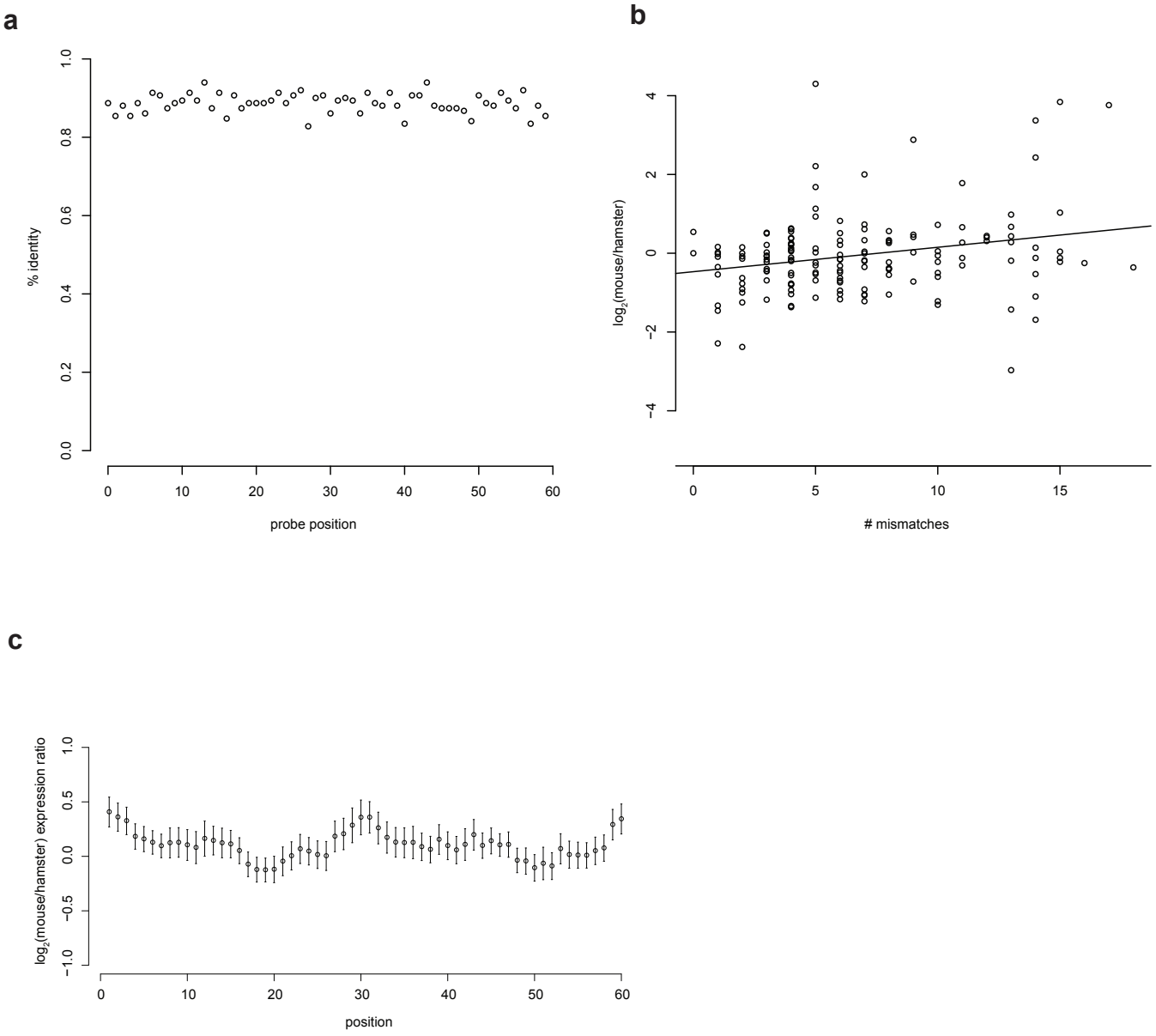
Supplementary Figure 1. CGH and expression analyses. **(a)** CGH analysis of diploid hamster kidney DNA compared to recipient hamster A23 cell DNA (chromosome 3). There are no regions of copy number increase or decrease in the A23 cells. **(b)** Bimodal distribution of CGH intensities in RH panel. **(c)** Chromosome 8 shows slight increases in retention in the T31 RH panel at the centromere and telomere. **(d)** Clustergram showing an overview of the $\log_{10}$(RH/A23) expression ratios for the radiation hybrid (RH) clones and the A23 recipient cells. The rows represent the expression data and the columns represent the RH clones and A23 cells (second to right column) ordered by similarity.

Figure S2



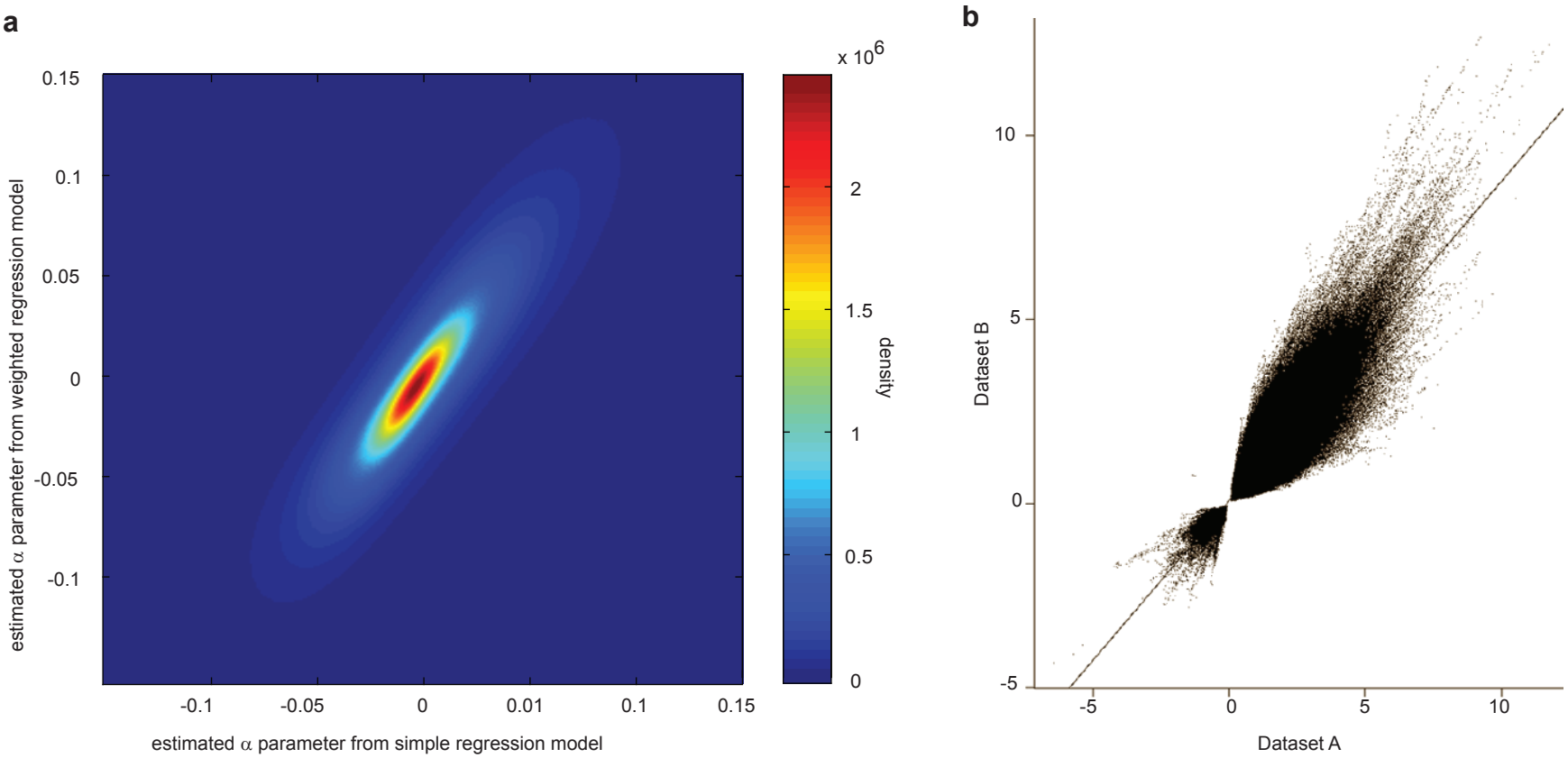
Supplementary Figure 2. Detection of mouse and hamster transcripts. **(a)** Scatterplot relating mouse and hamster brain $\log_2$(transcript levels) averaged over two arrays (R = 0.867, $P < 2.2 \times 10^{-16}$). **(b)** Scatterplot relating mouse and hamster liver $\log_2$(transcript levels) averaged over two arrays (R = 0.868, $P < 2.2 \times 10^{-16}$). **(c)** Scatterplot relating mouse and hamster $\log_2$(transcript levels) averaged over brain and liver (R = 0.863, $P < 2.2 \times 10^{-16}$). **(d)** Distribution of $\log_2$(mouse/hamster) expression ratios averaged over brain and liver.

Figure S3



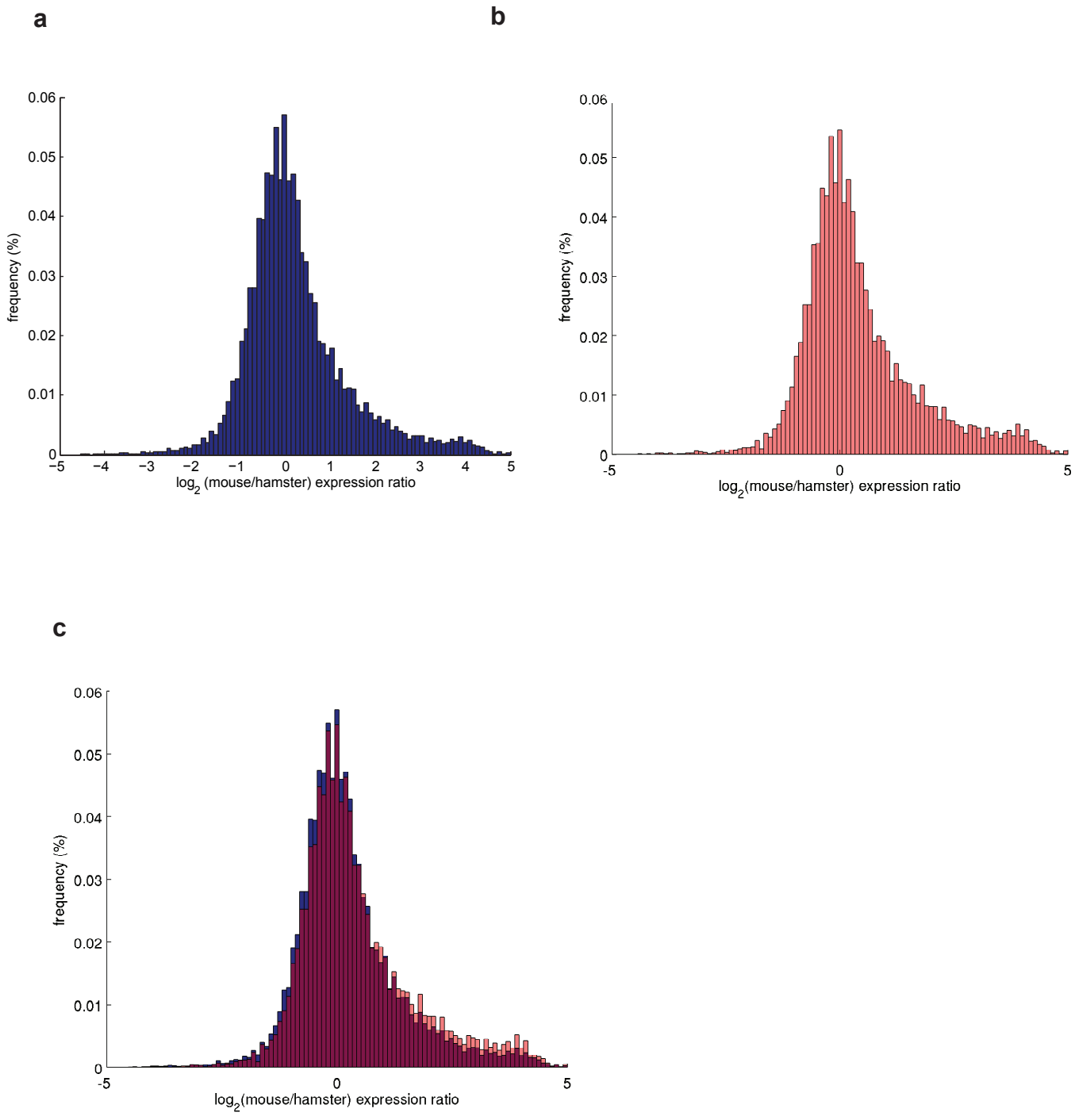
Supplementary Figure 3. Mouse/hamster sequence conservation. **(a)** Sequence conservation for 152 overlaps of hamster sequences with the oligonucleotides on the mouse Agilent expression array. The 60 nucleotides of the oligonucleotides are shown 5' to 3'. **(b)** Regression between the average $\log_2$(mouse/hamster) expression ratio for each of the 152 probes and the number of mismatches between the two species (R = 0.25, P = 0.002). **(c)** Relation between position of mouse/hamster mismatch on the 60-mer and the $\log_2$(mouse/hamster) expression ratio. The $\log_2$(mouse/hamster) expression ratio data is averaged over a sliding window of 5 nucleotides.

Figure S4



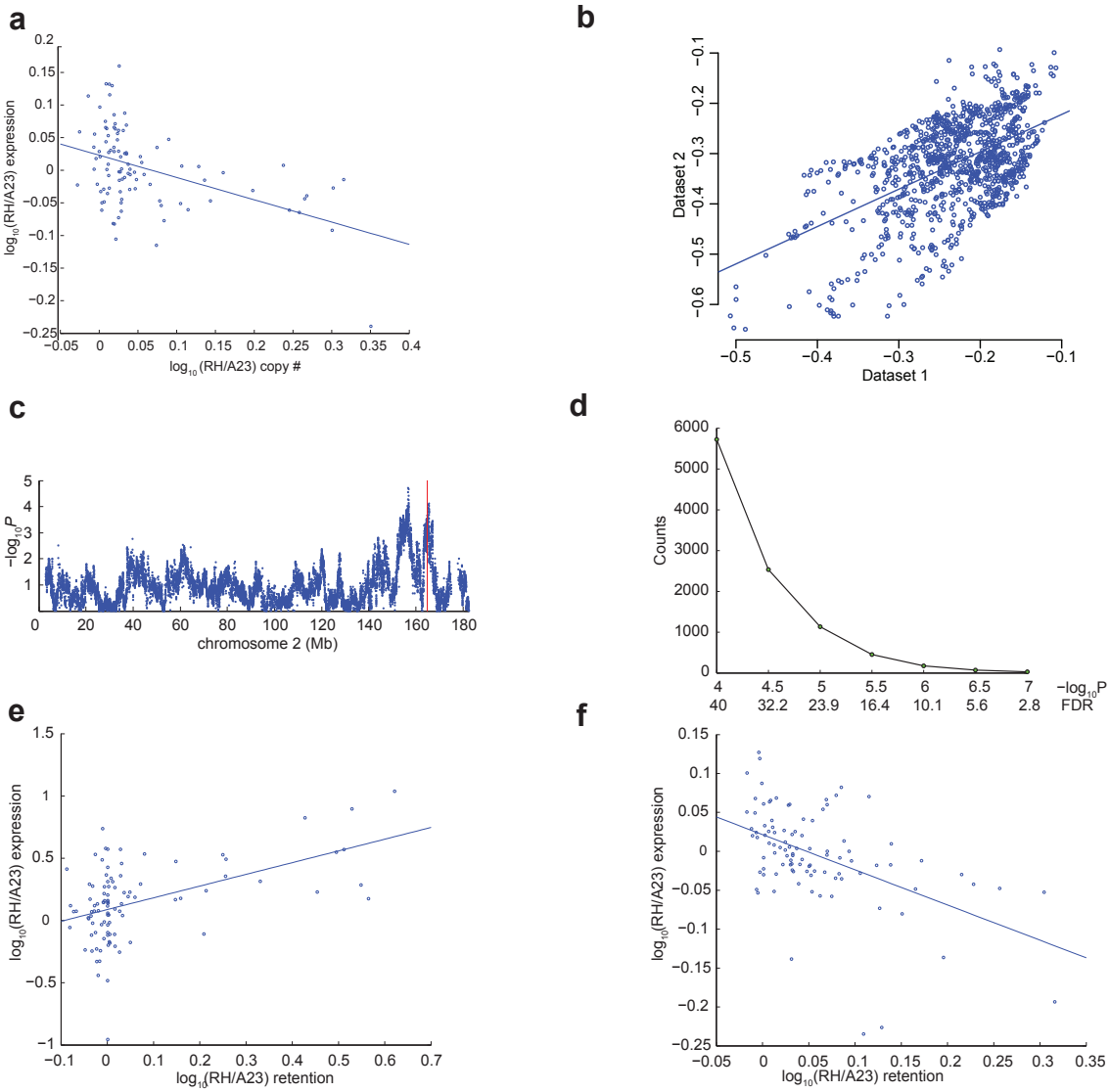
Supplementary Figure 4. Weighted regression model and replicate datasets. **(a)** Distribution of the estimated α parameters from the simple regression model (Model 1) and the weighted regression model (R = 0.8934, $P < 10^{-300}$). Outlier CGH markers with retention frequency < 5% or > 95% were excluded. Non-normalized CGH data were used for the simple regression model to make the estimated α parameters comparable to those from the weighted regression model on the same scale. The color bar indicates the density of points in the distribution. **(b)** Regression between the α values of the two replicate datasets using Model 1 for all FDRs < 0.4 (R = 0.887, $P < 2.2 \times 10^{-16}$).

Figure S5



Supplementary Figure 5. Distribution of average $\log_2$(mouse/hamster) expression ratios for cis and trans ceQTLs. **(a)** Distribution for cis ceQTLs **(b)** Distribution for trans ceQTLs. **(c)** Combined distribution. Pink bars show ratios with more trans than cis ceQTLs, dark blue with more cis than trans ceQTLs and magenta the overlap.

Figure S6



Supplementary Figure 6. Negative α cis ceQTL and trans ceQTLs lacking genes. **(a)** Regression between CGH copy number signal and gene expression for a negative α cis ceQTL showing outliers do not play a role. Cis ceQTL is for the WAP four-disulfide core domain 15A (*Wfdc15a*) gene on chromosome 2 at 164 Mb. α = −0.342, $-\log_{10}P$ = 4.709. **(b)** Regression between the α values of the negative α cis ceQTL for the two replicate datasets with FDRs < 0.4 (R = 0.537, P = 2.2 x 10^{-16}). **(c)** $-\log_{10}P$ curve. **(d)** Relation between the number of trans ceQTLs lacking genes and their $-\log_{10}P$ and FDR values. The radius from the peak marker of the trans ceQTL was 300 kb. In this tally, trans ceQTLs regulating multiple genes (hotspots) were counted multiple times. **(e)** Regression between gene expression and CGH copy number ratio at the peak marker for a trans ceQTL with positive α lacking genes, located on chromosome 1 at 49 Mb regulating the medium-wave-sensitive opsin 1 cone pigment (*Opn1mw*) gene on the X chromosome. α = 0.942, $-\log_{10}P$ = 5.104. **(f)** Regression for a trans ceQTL with negative α lacking genes, located on chromosome 6 at 74 Mb regulating the *AI464729* gene on chromosome 16. α = −0.452, $-\log_{10}P$ = 5.390.