

Intra- and inter-individual genetic differences in gene expression

Mark J. Cowley · Chris J. Cotsapas · Rohan B. H. Williams ·
Eva K. F. Chan · Jeremy N. Pulvers · Michael Y. Liu ·
Oscar J. Luo · David J. Nott · Peter F. R. Little

Received: 4 November 2008 / Accepted: 23 February 2009 / Published online: 8 May 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract Genetic variation is known to influence the amount of mRNA produced by a gene. Because molecular machines control mRNA levels of multiple genes, we expect genetic variation in components of these machines would influence multiple genes in a similar fashion. We show that this assumption is correct by using correlation of mRNA levels measured from multiple tissues in mouse

strain panels to detect shared genetic influences. These correlating groups of genes (CGGs) have collective properties that on average account for 52–79% of the variability of their constituent genes and can contain genes that encode functionally related proteins. We show that the genetic influences are essentially tissue-specific and, consequently, the same genetic variations in one animal may upregulate a CGG in one tissue but downregulate the CGG in a second tissue. We further show similarly paradoxical behaviour of CGGs within the same tissues of different individuals. Thus, this class of genetic variation can result in complex inter- and intraindividual differences. This will create substantial challenges in humans, where multiple tissues are not readily available.

M. J. Cowley, C. J. Cotsapas, and R. B. H. Williams contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00335-009-9181-x) contains supplementary material, which is available to authorized users.

M. J. Cowley · C. J. Cotsapas · R. B. H. Williams ·
E. K. F. Chan · J. N. Pulvers · M. Y. Liu ·
O. J. Luo · P. F. R. Little (✉)
School of Biotechnology and Biomolecular Sciences,
The University of New South Wales, Sydney,
NSW, Australia
e-mail: peter.little@nus.edu.sg

M. J. Cowley
The Garvan Institute of Medical Research, Darlinghurst,
NSW, Australia

C. J. Cotsapas
Centre for Human Genetic Research, Massachusetts General
Hospital, Boston, MA, USA

C. J. Cotsapas
Harvard Medical School and The Broad Institute
of MIT and Harvard, Cambridge, MA, USA

R. B. H. Williams · P. F. R. Little
The Clive and Vera Ramaciotti Centre for Gene Function
Analysis, The University of New South Wales, Sydney,
NSW, Australia

R. B. H. Williams
John Curtin School of Medical Research,
Australian National University, Canberra, Australia
e-mail: rohan.williams@anu.edu.au

E. K. F. Chan
Department of Plant Sciences, University of California
Davis, Davis, CA, USA

D. J. Nott
School of Mathematics and Statistics,
The University of New South Wales, Sydney,
NSW, Australia

D. J. Nott
Department of Applied Probability and Statistics,
Singapore, Singapore

P. F. R. Little
National University of Singapore, Singapore,
Singapore

Introduction

Gene expression is controlled by molecular machines that each interact with multiple genes and transcripts. It follows that genetic variation in the components of these machines could, in principle, simultaneously alter the final level of mRNA derived from multiple genes. In this study we set out to identify groups of mouse genes whose mRNA levels are simultaneously influenced by genetic variation. We report shared influences of genetic variation on the mRNA levels of large numbers of genes but also show that shared group behaviour can vary unpredictably between different tissues of the same individual and between the same tissues of different individuals.

We use the term regulatory variation (RV) to describe any genetic variation that affects the amount of mRNA produced from a gene; it can occur through the disruption of *cis*-regulatory sequences, such as promoter or enhancer elements, or more importantly for our observations, through changes to *trans*-acting components of the molecular machines that control the amount of steady-state mRNA in a cell, such as transcription or splicing complexes. Recent studies have shown that these processes are subject to significant influences of genetic variation that result in heritable changes to final mRNA levels (Cotsapas et al. 2006; Gibson and Weir 2005; Rockman and Kruglyak 2006; Williams et al. 2007). The majority of findings to date, using predominantly expression quantitative trait loci (eQTL) experimental designs, suggests that *cis*-acting RV appears to have a larger-effect size and is thus more easily detected; in comparison, *trans*-regulatory variation appears to have a smaller-effect size and is either less common or harder to detect (Goring et al. 2007; Petretto et al. 2006; Stranger et al. 2005). When *trans*-acting influences are identified, there tends to be a small number of eQTLs that influence the expression of large numbers of genes, so called “master-regulators” of gene expression, suggesting that RV affects the expression level of groups of genes simultaneously (Keurentjes et al. 2007; Mehrabian et al. 2005; West et al. 2007; Wu et al. 2008).

However, while eQTL analysis is an appropriate approach to investigate the effects caused by a small number of genetic influences, each with large-effect sizes, it has limited or no power to detect multiple, small-effect eQTLs that likely would be the biological basis of RV in the *trans*-acting molecular machines (Brem and Kruglyak 2005; Williams et al. 2007). To help overcome this limitation, several groups have used correlation-based approaches to identify groups of genes that covary under the influence of simple or complex genetic influences (Brem and Kruglyak 2005; Chen et al. 2008; Emilsson et al. 2008; Ghazalpour et al. 2006; Keller et al. 2008; Williams et al. 2007). The conceptual basis of such experiments is simple:

mRNA levels that vary similarly across multiple individuals are likely to do so because of shared sensitivity to genetic influences. Correlation-based approaches focus on detecting RVs by their shared outcome at the expense of attempting to identify the individual RVs themselves.

In this study of inbred and recombinant inbred mice, we set out to investigate *trans*-acting RV using correlation analysis to identify groups of genes that are likely to be influenced by shared RV, and thus shared regulatory factors, and to investigate the consequence of *trans*-acting RV in three different mouse tissues to assess the degree to which the outcome of such RV is the same in all tissues. We use correlation-based methods to show that the effects of RV are, as predicted, coordinated changes to the mRNA levels of groups of genes. These group changes can be very different in multiple tissues of the same individual and different in the same tissues from multiple individuals.

To identify genes whose expression levels may be affected by RV and to investigate their regulation in multiple tissues, we adopt the following experimental design: First, we compare gene expression levels in three tissues of two inbred mouse strains, C57BL/6J and DBA/2J, and of 31 strains of the BXD recombinant inbred (RI) panel derived from these two progenitors. Next, we look for genes whose expression differs between the progenitor strains in at least one of these tissues; within these we identify subsets of genes whose mRNA levels vary coordinately across the BXD RI strains and the three tissues. We call these “correlating groups of genes” or CGGs (see the subsection “[Identifying groups of genes under shared genetic influence in multiple tissues](#)”). We then validate the shared regulatory influences acting upon these CGGs by testing the conservation of their expression changes in both the parental strains and in the distantly related inbred strain SJL/J (see the subsection “[Intra- and interstrain expression variation in CGGs](#)”) and explore the biological functionality and putative transcriptional architecture associated with the CGGs (see the subsection “[CGG and biological functions](#)”).

Materials and methods

RNA preparation

Eight-week-old male *Mus musculus* strains C57BL/6J, DBA/2J, and SJL/J were obtained from the Biological Resources Centre, UNSW (Sydney, Australia), and *Mus musculus* BXD/TyJ strains 1, 2, 5, 6, 8, 9, 11–16, 18–24, 27–34, 36, 38–40, and 42 were obtained from the Jackson Laboratory (Bar Harbor, ME, USA). Following cervical dislocation, whole brain, kidney, and liver tissues were harvested according to protocols approved by the University

of New South Wales Animal Care and Ethics Committee (Ethics Code ACEC 01/43) and snap-frozen in liquid N₂. Total RNA was extracted according to the manufacturer's instructions with TRIzol Reagent (Invitrogen, Mt. Waverley, Victoria, Australia); purity and integrity were assessed by OD₂₆₀/OD₂₈₀ readings greater than 2 and intact rRNA bands (Agilent Bioanalyzer, Agilent, Forest Hills, Victoria, Australia), respectively.

DBA/2J vs. C57BL/6J strain experiment Total RNA from the three tissues of ten individuals was pooled for each strain (9 for liver) to remove individual variation in gene expression; 20 µg of pooled RNA and 2 µg of Lucidea Universal Scorecard Spike-in (Amersham Biosciences, Castle Hill, NSW, Australia) were reverse transcribed using the SuperScript III Indirect cDNA Labelling System (Invitrogen) and fluorescently labeled with Alexa Fluor 555 for C57BL/6J and Alexa Fluor 647 for DBA/2J (Invitrogen).

BXD panel experiments Equal amounts of total RNA from three animals from each BXD strain were mixed to yield tissue pools representative of the genetic backgrounds. A common reference sample was created for each tissue from total RNA extracted from ten 8-week-old male C57BL/6J mice (a different RNA source than the parental strain experiment). Twenty micrograms of pooled RNA was reverse transcribed (as above) and fluorescently labeled with Alexa Fluor 555 for C57BL/6J and Alexa Fluor 647 for BXD strain samples (as above).

C57BL/6J vs. SJL/J experiment Total RNA from the brain, kidney, and liver of five C57BL/6J and five SJL/J individuals was pooled for each strain. cDNA synthesis was same as for the C57BL/6J vs. DBA/2J experiment, but sodium tetraborate instead of sodium bicarbonate was used in the labeling buffer. Again, C57BL/6J cDNA was labeled with Alexa Fluor 555 and SJL/J with Alexa Fluor 647 for DBA/2J (Invitrogen).

Microarray experiments

DBA/2J vs. C57BL/6J experiment For each tissue, labeled cDNA was directly compared on six replicate glass-slide two-colour microarrays containing the Compugen Mouse OligoLibrary representing 21,997 genes and Lucidea Universal ScoreCard (Clive and Vera Ramaciotti Centre for Gene Function Analysis, UNSW, Sydney, Australia), in 100 µl of DIGEasy buffer (Roche, Basel, Switzerland) with 5 µl each yeast tRNA and calf thymus DNA as blockers (Invitrogen). Utility controls from the Lucidea Scorecard were not used and therefore served as additional negative controls. Hybridised microarrays were washed in 1 × SSC, three times in 1 × SSC, 0.1% SDS at 50°C, and three times in 1 × SSC, dried by centrifugation, and scanned with the

GenePix 4000B microarray scanner (Axon Instruments, Union City, CA, USA).

BXD panel experiments Identical arrays and processing were as above, with one array being performed for each tissue in each BXD strain, giving a total of $31 \times 3 = 93$ arrays.

C57BL/6J vs. SJL/J experiment Identical arrays and processing were as above, but three microarrays per tissue were performed per tissue, giving a total of $3 \times 3 = 9$ arrays.

Data processing

Image analysis was performed with the Spot image analysis software ver. 2 (CSIRO, Australia, <http://experimental.act.cmis.csiro.au/Spot/index.php>). All further data processing and statistical analyses were performed using R ver. 2.0.0–2.6.1 (Ihaka and Gentleman 1996). Gene expression data were morph background corrected and log₂ transformed. Data for controls and the 232 replicated spots of the housekeeping gene *Gapd* (NM_008084) were removed prior to normalization to avoid bias.

DBA/2J vs. C57BL/6J experiment All 18 slides were then normalized for intensity and spatial bias using print-tip loess and then quantile adjusted to adjust for the differing scale of measurements across arrays (Yang et al. 2001). Replicate slides were averaged.

BXD panel experiments All 93 slides were normalized using print-tip loess. To standardise across experiments from the three tissues, we subselected the data from genes considered to be expressed in all three tissues in the parental experiment and then applied quantile normalization. The log₂ ratios of intensities, $M = \log_2 R - \log_2 G$ (referred to as *M* values), were subsequently used as expression measurements.

C57BL/2J vs. SJL/J experiment Processing was that as for the parental experiment.

Differential expression in parental strains across multiple tissues

We classified genes as reliably detected if their log mean intensity, $A = 0.5 \times (\log_2 R + \log_2 G)$, in all three tissues was greater than the 95th percentile of negative controls present on our arrays. *B* statistics were then calculated for all genes using default parameters in the R limma library ver. 1.8.6 (Smyth 2004), part of the Bioconductor project (Gentleman et al. 2004). Genes were classified as genetically influenced if they had both a *B* statistic (LOD) greater than 3 and an *A* value greater than the intensity threshold. In all three tissues there were 6,075 genes detected above threshold. Of these, 755 were genetically influenced in one or more tissues.

Cross-tissue correlation analysis

To identify the genes that have expression patterns similar to those of gene g_i in all tissues, we adopted a correlation-based approach. There are three per-tissue expression matrices, E_{brain} , E_{kidney} , and E_{liver} , each of dimension $G \times S$, where G is the number of genes and S is the number of strains, i.e., 755 genes \times 31 strains in the present case. Pairs of genes that are correlated with each other in all three tissues are of primary interest because they may be under the influence of some common, tissue-independent regulatory mechanisms. We identify such pairs of genes by joining the three per-tissue expression matrices E_{brain} , E_{kidney} , and E_{liver} into a single $G \times 3S$ cross-tissue expression data matrix $E_{\text{BKL}} = (E_{\text{brain}}|E_{\text{kidney}}|E_{\text{liver}})$. We then computed a $G \times G$ correlation matrix, C_{BKL} , from E_{BKL} using Spearman's ρ as a correlation metric. C_{BKL} is referred to as the *cross-tissue correlation matrix*. C_{BKL} was then hard thresholded for various values of $|\rho|$, thus defining the adjacency matrix C^*_{BKL} , which represents an undirected simple network. In the present study, all networks were generated using a threshold of $|\rho| \geq 0.775$. Statistical analyses related to threshold selection are provided in the Supplementary Results and Supplementary Fig. 1. Nodes in the cross-tissue correlation matrix were laid out using the 2D Fruchterman–Reingold algorithm (Fruchterman and Reingold 1991) and visualised using implementations available in the *igraph* library in R/Bioconductor (Gentleman et al. 2004).

CGG centroid R^2 analysis

The centroid of each CGG is the per-strain average M value for all genes in the CGG, which we calculated for each tissue independently or from all three tissues combined. To

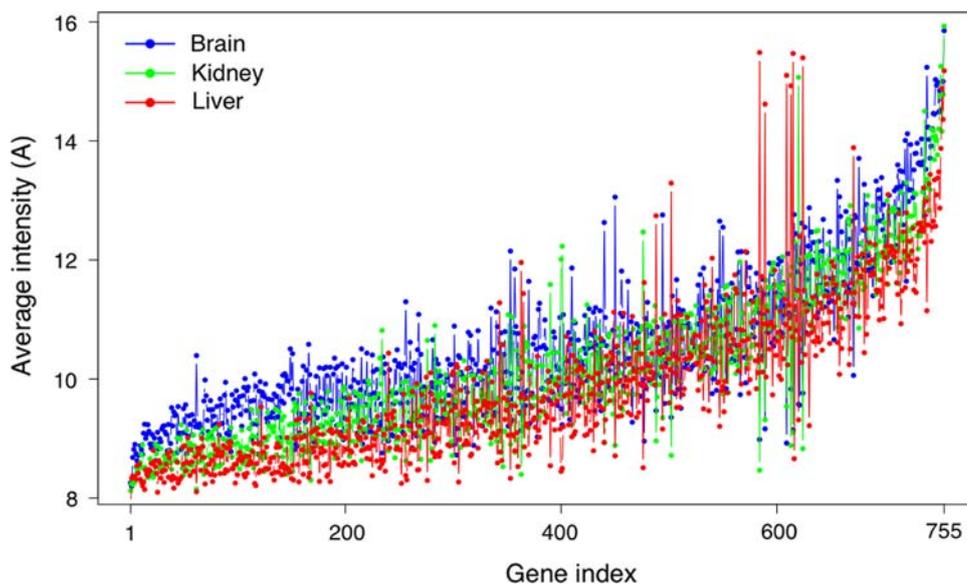
determine the similarity of each gene in the CGG to its centroid, we compute R^2 as the square of Pearson's product–moment coefficient (r), obtaining a distribution of R^2 values for all genes in the CGG. We assess the statistical significance of the observed R^2 using permutation analysis. We repeat this analysis for random CGGs, chosen by randomly sampling the same number of genes from the set of 755 genes, and obtain a distribution of R^2 values for each gene in the random CGG to the random CGGs centroid. We compare the observed distribution of R^2 to the random distribution using the Mann–Whitney U test using the upper-tail P values. We repeat this for 1,000 random CGGs and count the number of times the P value was less than 0.05, divided by the number of permutations. Similar results were obtained if the random genes were resampled from the set of 6,075 genes, or if the random genes were compared to the observed CGGs centroid rather than the random CGGs centroid (data not presented).

Coherency test statistic

The coherency test statistic is designed to measure how consistent the directionality of relative expression is in a set of genes (see Results subsection “[Intra- and interstrain expression variation in CGGs](#)”). Given the expression ratios (M values) from the comparison of two strains (such as C57BL/6J vs. DBA/2J) and a set of genes, $G = \{g_1, \dots, g_N\}$, with corresponding measurements of average relative expression, \hat{M}_g , across a set of replicates associated with each gene, the vertex-based coherency C_G is calculated as follows:

$$C_G = \frac{\sum_{k=1}^N \text{sign}(\hat{M}_{g_k})}{N}$$

Fig. 1 The average expression level (A value) in brain (blue), kidney (green), and liver (red) for each of the 755 genetically influenced genes in DBA/2J and C57BL/6J is plotted, with expression level on the y axis (\log_2 scale), ordered from left to right with increasing average expression in all three tissues. Note the deliberate absence of genes that are expressed in only one of the three tissues due to limiting to those genes that were expressed above background in all three tissues (see [Materials and methods](#))



where sign is the sign function which is defined as

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

Thus, this vertex-based coherency score is in the range $[-1, 1]$, with values that are closer to either extreme indicating more coherently downregulated (-1) or upregulated ($+1$) expression. For example, for a group of ten genes, if nine are upregulated and one is downregulated, then the coherency is $(+9 - 1)/10 = 0.8$.

Permutation test We chose 1,000 random sets of G genes from a set of 755 genes (by permuting gene labels) and assessed the significance of the observed coherency of each CGG using the following formula:

$$P = \frac{\#\{|C_G^*| \geq |C_G|\}}{B}$$

where G^* denotes a randomised version of gene set G , defined using the label-permuted set of 755 genes, and B is the number of such permutations generated. For example, if the given CGG had a vertex-based coherency score of 0.77, and of 1,000 randomised samples only six scores were observed to be greater than 0.77, then the P value would be $6/1,000 = 0.006$. Further material relating to the development and validation of the coherency test statistic is provided as Supplementary Material.

Gene Ontology (GO) analysis

To test for enrichment or depletion of a GO term in a set of genes of interest, we tested whether genes of interest were mapped to the GO term at a level greater than chance expectation (defined as the observable proportion of genes mapping to the term in the set of expressed genes in the experiment) using sampling without replacement from the hypergeometric distribution (using the *phyper* function in R). We used a strict Bonferroni correction for $P < 0.05$, corrected for the number of terms with more than five genes annotated to them, either directly or via transitive relationships in the ontology. We used the Bioconductor package *GO* (v1.1.14) and mapped microarray identifiers (GenBank IDs) to Entrez Gene IDs based on probe-sequence-similarity using custom scripts (available on request).

Analysis of genomic location

Using data available from the UCSC Genome Bioinformatics database (www.genome.ucsc.edu), we mapped the genomic location of genes in CGG 1 through CGG 5 and ordered them according their physical location across the genome (mm9). We then computed the distance between

each adjacent gene and examined instances where adjacent genes were located within 1 Mb of each other, considering two genes colocalised if they occurred within 1 Mb of each other.

Transcription factor binding motifs

The GenBank sequences for each of the 6,075 expressed genes were aligned to the NCBI 35.1 build of the mouse genome using BLAT (v32x1) (Kuhn et al. 2007), and the best hits were retained. The upstream 1,000 bp from these sequences were then retrieved using BioPerl and converted into FastA formatted files. Repeat regions were masked to lower-case letters using RepeatMasker (ver. open-3.1.6) and RepBase (ver. 20061006) using the following flags: “-species mouse -xsmall -gff”. Then the upstream sequences for all of the genes in each CGG were separated into a separate FastA formatted file. The Transcription Factor motif library from JASPAR (Vlieghe et al. 2006) was downloaded (jaspar2005core) and formatted to suit CLOVER using tools from the CLOVER download page (<http://zlab.bu.edu/clover>) (Frith et al. 2004). CLOVER [Cis-eLement OVERrepresentation (ver. March 29, 2006)] was run to search for overrepresented motifs in the upstream sequences from the genes in each CGG compared to a background set of sequences from the 6,075 expressed genes. These data were permuted 1,000 times to generate P values for over-/underrepresentation in the data sets. The following flags were used when running CLOVER: “-1 -t 0.05”. We subsequently used the AACAAAT motif to represent the entire family of *Sox* transcription factors (Koopman 2001).

eQTL analysis for genes in CGG 2

For all expression phenotypes in CGG 2 (63 genes), we calculated linkage test statistics for the closest genetic marker (www.webqtl.org) (Chesler et al. 2004) to each of the 24 transcription factor encoding genes whose binding motifs were enriched in the proximal promoters of genes in CGG 2. This analysis was performed in each of the three tissues separately. We estimated the significance of linkage to each marker using likelihood ratio statistics (LRS) and model-based P values calculated using the QTL Reaper code (v1.1.0 with single-marker analysis option; www.genenetwork.org/qtlreaper.html). We corrected the number of comparisons (marker \times gene \times tissue) using the Bonferroni correction. We performed a genome-wide linkage analysis using the centroids of the each CGG (see the subsection “CGG centroid R^2 analysis”) as an expression trait and corrected the number of comparisons (marker \times tissue) using the Bonferroni correction.

Results

Identifying groups of genes under shared genetic influence in multiple tissues

Using microarrays, we began by identifying genes that were differentially expressed in at least one of the three tissues, whole brain, kidney, or liver, between strains C57BL/6J and DBA/2J. We found that we could reliably detect 6,075 transcripts above background in all three tissues, of which 755 were variably expressed between the two strains at a $\text{LOD} > 3$ in any of the three tissues, using the B statistic of Lönnstedt (Lönnstedt and Speed 2002), as modified by Smyth (2004) (see Methods subsection “[Differential expression in parental strains across multiple tissues](#)”). We ascribe this consistent variation in gene expression to regulatory variation, since environmental factors have been reduced to a minimum. We stress that we have deliberately avoided including genes that are expressed in a “tissue-specific” manner, in the sense of being expressed in only one or two of the three tissues (Fig. 1).

The identification of 755 genes as potential targets of regulatory variation(s) does not allow us to find out if each gene is under a unique or shared influence. To do this, we need to study the 755 genes in multiple, changing, genetic backgrounds, reasoning that we could then detect shared influence by detecting correlated alterations of mRNA levels. Such correlated changes could in principle be observed between genes within either single or multiple tissues. We chose to search for mRNA correlations across multiple tissues in the first instance and then further studied the behaviour in the individual tissues, seeking to find if the outcome of genetic influence on genes is the same in each tissue.

To achieve this, we measured mRNA levels of the 755 genes in the same three tissues in 31 BXD recombinant inbred (RI) strains (Taylor et al. 1999), pooling three age- and sex-matched mice from each (see Methods subsection “[RNA preparation](#)”). Although appropriate for the correlation-based analysis performed here (see below), we note that pooling removes any information about within-strain variability and thus limits estimation of transcript heritability (e.g., Petretto et al. 2006). These strains have been derived from crosses of C57BL/6J and DBA/2J, which have been bred to homozygosity by repeated sibling pair mating. Because they carry arbitrary mixtures of the two progenitor backgrounds but are homozygous at each locus, it follows that most strains will have inherited some of the C57BL/6J alleles and some of the DBA/2J alleles of any factors, basal or conditional, controlling the mRNA levels of the 755 genes. If these factors influence more than a single transcript, we would predict that the levels of these co-influenced mRNAs would correlate across the BXD panel, thus forming a CGG.

To identify those genes that have similar expression patterns in all 31 BXD strains and in all three tissues, we adopted a correlation-network approach, permitting the summarisation of a large amount of complex data into a form that is easily visualised and interpreted (Freeman et al. 2007) (see Methods subsection “[Cross-tissue correlation analysis](#)”). Rather than construct a correlation network for each tissue and identify groups of genes that are found in all three networks, we first combine the gene expression data for all three tissues together and then construct a cross-tissue correlation network. This approach is advantageous in that we can identify relationships between pairs of genes that may be correlated more weakly in one of the tissues yet strongly correlated in the other two.

We initially performed single linkage to gain overall insight into the interrelationships between genes in the cross-tissue matrix, revealing the existence of several major subgroupings of genes (Fig. 2). We chose to construct networks using the well-established, widely used approach of thresholding correlation matrices (Butte and Kohane 2000; Freeman et al. 2007; Gower and Ross 1969; Voy et al. 2006; Wirth et al. 1966). As our primary aim was to identify groups of coregulated genes that are plausibly under common genetic control, we focused on finding groups of interconnected genes that are distinct from other such groups (referred to as *connected components* in graph theoretical terms) (Freeman et al. 2007; Wirth et al. 1966). An important step in identifying such groups of genes is choice of threshold on the correlation matrix: too low a threshold will result in a densely interconnected network, while too high a threshold will result in a sparsely populated, loosely interconnected network (Freeman et al. 2007). We performed statistical analyses to gain insight into the suitability of a range of thresholds and constructed our cross-tissue correlation network using a correlation threshold of $|\rho| = 0.775$ (see Methods subsection “[Cross-tissue correlation analysis](#)” and supplementary results for full details).

This cross-tissue correlation network contained 212 (28.1%) genes that correlate with at least one other transcript; the genes have a median degree of 4, with 73% of genes having a degree of 2 or more (Fig. 3a). These genes are central to our subsequent study; in principle, they are influenced by regulatory variation(s) that influence mRNA levels in all three tissues simply because the correlation statistic is calculated across all three tissues. Performing similar analyses on subsets of tissues, we find that at the same threshold an additional 204 (27.0%) genes are correlated in any pair of tissues, and an additional 191 (25.3%) are correlated in any single tissue. A total of 607 (80.4%) of the 755 genes exhibit correlated behaviour in any network, suggesting that shared regulatory influences upon gene expression are widespread, and over 55% are correlated in multiple tissues (data not presented).

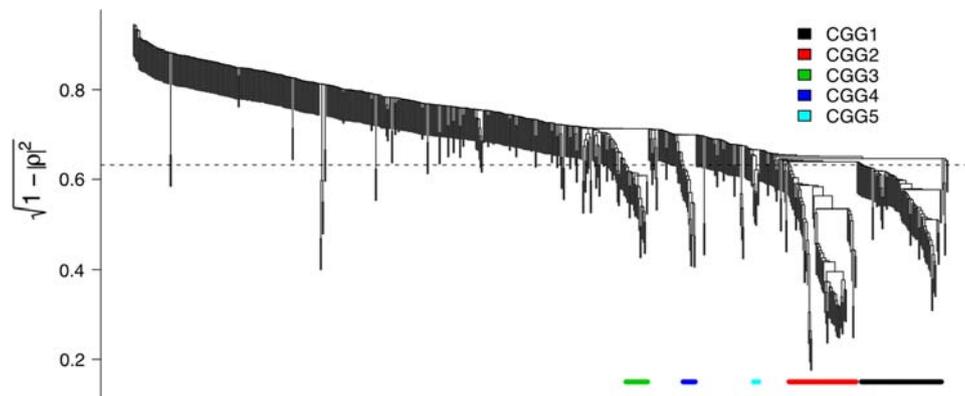


Fig. 2 Dendrogram formed by single-linkage hierarchical clustering of the 755-gene cross-tissue matrix. The distance metric used was a monotonically decreasing function of the absolute value of Spearman's ρ , namely, $\sqrt{1 - |\rho|^2}$. The value of this metric equivalent to

that used to generate the correlated groups of genes (CGGs) ($0.623 \equiv |\rho| = 0.775$) is shown as a horizontal dashed line. Genes in the five largest CGGs are colour-coded

In the original cross-tissue network, we find that the 212 genes fall into 19 discrete connected components, which we shall now refer to as correlating groups of genes or CGGs. Of these groups, ten contain at least three genes and the largest five contain 75, 63, 21, 12, and 6 genes, respectively. The cross-tissue network is displayed in Fig. 3a along with the expression patterns across the 31 BXD strains and the three tissues; it is displayed for the five largest CGGs in Fig. 3b.

We calculated the amount of variation in a gene's mRNA level that could be accounted for by the shared behaviour of the CGG. To do this, we used the coefficient of determination (R^2) to compare the expression pattern of each transcript to the centroid of their respective CGG (grey and thick coloured lines in Fig. 3b, respectively) for each tissue individually and across all three tissues simultaneously (see Methods subsection “CGG centroid R^2 analysis”). Collectively, these R^2 values ranged from 0.00 to 0.95, and the mean variability explained by the CGG centroid in each tissue (or in the combination of three tissues) ranged from 0.52 to 0.79 (with the exception of CGG 1 in brain which demonstrated low R^2 values) (Fig. 4). That is, on average, more than 50% of variation in a gene's mRNA level can be accounted for by the shared influences of regulatory variation on the CGG (see also supplementary results and Supplementary Fig. 2 for additional results relating to the statistical robustness of these observations). To place this result into context, Petretto et al. (2006) suggest that mappable *cis*-acting eQTL can account for 31–51% of the heritability of a mRNA level and mappable *trans*-acting eQTL can account for 14–21%.

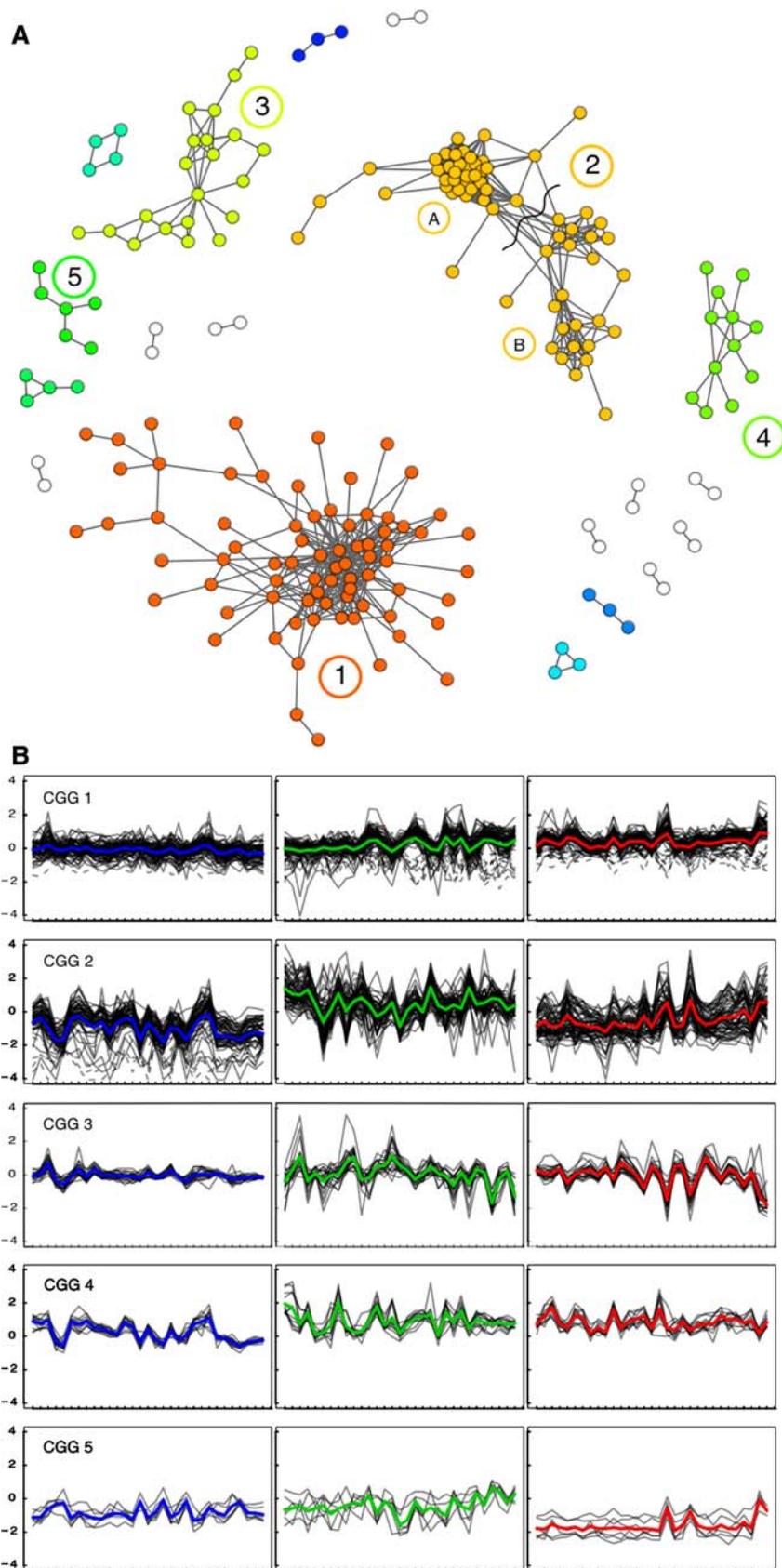
While we have illustrated the congruous behaviour of mRNAs within a CGG, we also note from Fig. 3b that mRNA level profiles are strikingly different between each tissue. This is supported by calculating the correlation between the intratissue centroids for each CGG

(Supplementary Table 1): The only statistically significant relationship is, in fact, an anticorrelation between the centroids of CGG 2 in brain and liver ($\rho = -0.59$, $P = 5.94 \times 10^{-4}$). These results show that while genes within a CGG are highly correlated to each other, consistent with the idea of being influenced by shared factors, the outcome of such regulation is markedly different in each tissue such that the overall pattern of a group's expression in each tissue is at best uncorrelated or even anticorrelated. These differences can be explained by regulatory variation residing in either multiple regulatory components that act in a tissue-specific fashion or in a single cross-tissue component whose activity or expression is itself modulated by tissue-specific factors.

Intra- and interstrain expression variation in CGGs

Having identified CGGs based on their expression patterns in three tissues across a panel of BXD mice, we sought independent evidence that the expression of these groups of genes are being influenced in a coordinated fashion, due to the effects of genetic as opposed to other sources of variation. Within each individual BXD animal, all genes in a CGG should be coordinately regulated, even if this differs across tissues. If these levels are indeed due to genetic differences in the regulatory factors controlling the ultimate mRNA level, then we would expect that CGG members should display similar correlated expression patterns across different genetic backgrounds. However, the multiple, complex changes in genetic background implicit in this experiment are unlikely to result in exactly the same mRNA levels in any two individuals; therefore, rather than test for the identical expression level of all genes in the CGG, we designed a test to detect for the identical direction of mRNA levels: relatively up- or downregulated, compared to a suitable reference or

Fig. 3 a Correlations between genes are displayed as a network: Edges connect two genes if those genes are correlated with an absolute value of Spearman's $\rho > 0.775$. Two hundred twelve of the 755 genetically influenced genes (see text) pass this threshold and are positioned in the x, y plane based on a 2D Fruchterman-Reingold layout algorithm (Fruchterman and Reingold 1991). Correlated groups of genes (CGGs) with at least three genes in them are coloured, and the five largest are numbered. When split into two parts, as per the black curved line, CGG 2 displays coherent expression patterns and functional clustering (see text). **b** Panels show the mRNA expression ratios of genes in the relevant CGGs measured in each BXD strain in three tissues (1st panel, brain; 2nd panel, kidney; and 3rd panel, liver). The vertical axis is the expression ratio vs. *C57BL/6J* (M values) of mRNA level in each of the 31 BXD strains (horizontal axis). Each individual gene's M values are plotted as black lines, with thick coloured lines representing the CGG centroids (blue, green, and red for brain, kidney, and liver, respectively). Note the striking differences in the expression patterns of the CGG centroids in the three different tissues



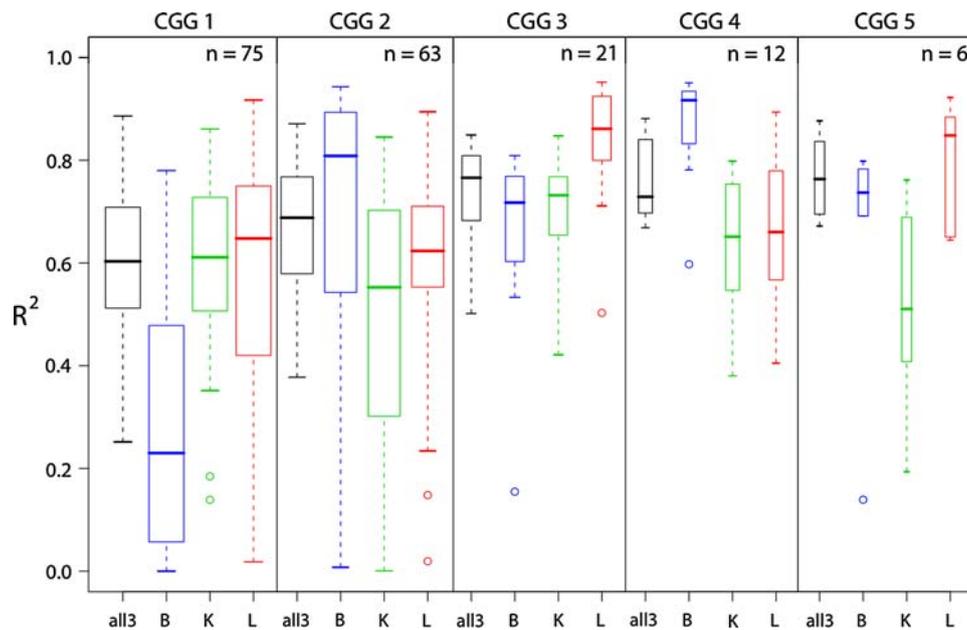


Fig. 4 The extent of shared rather than individual influence on a gene's expression level. Note that a very high proportion of the variation in many individual genes' mRNA levels can be accounted for by the influence of shared rather than gene-specific influences. Box-and-whisker plots represent the R^2 between each CGG's centroid to each gene in the CGG, independently evaluated for the brain mRNA levels only (B, blue), for the kidney (K, green), or for the liver

(L, red) or across all three tissues (all 3, black). Each box represents mRNA levels falling in the 25th to the 75th percentile, the thick line the median, and the whiskers extend to at most two standard deviations away from the median, with outliers indicated with circles. The horizontal width of each box is proportional to the number of genes in each CGG

baseline. This coordinated expression over all genes in a CGG can be summarised as a coherency statistic: the proportion of genes whose mRNA levels are upregulated (or downregulated) relative to the reference (see Fig. 5a for an overview, and Methods subsection "Coherency test statistic" for details). We performed simulation studies to assess the performance of the coherency statistic with respect to both the number of genes in a CGG and the magnitude and variability of the expression changes (see supplementary results). Simulating the conditions of our experiment, we identified that the score is adequately powered to detect coherent directionality of expression for CGGs of at least ten genes (at permuted $P < 0.05$). For groups of genes with less than ten genes, the score had little power, even in the case of maximal coherency.

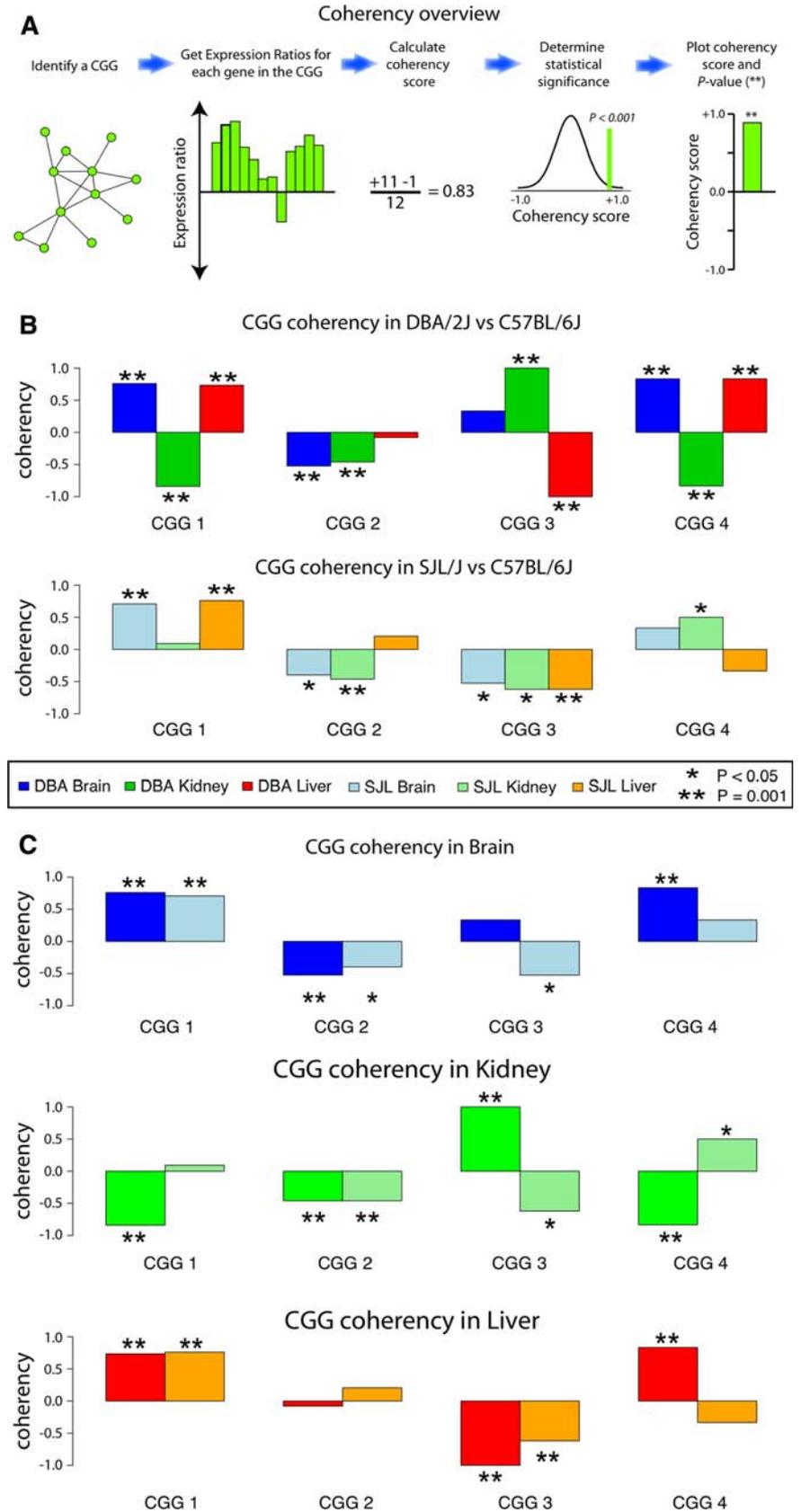
Given that we have hypothesised that each CGG is caused by genetic differences that existed between the parental strains DBA/2J and C57BL/6J, we looked at coherency in these two strains. For the four largest CGGs (those having between 75 and 12 genes), we found that all four were significantly coherent ($P = 0.001$) in at least one tissue (bars marked with ** in row 1 of Fig. 5b and Supplementary Table 2).

We note that CGG 1 in the brain, which had the lowest R^2 values to its centroid, nevertheless exhibits high coherency (coherency = 0.76; $P = 0.001$). While the shared

contribution to overall mRNA levels of the CGG might be relatively small, there is a marked effect upon the direction of mRNA level changes. We also note that the 63 genes in CGG 2 have complex properties: coherency is moderate, but still significant in brain (coherency = -0.52 ; $P = 0.001$) and kidney (coherency = -0.46 ; $P = 0.001$) and not coherent in liver (coherency = -0.08 ; $P = 0.33$). However, close inspection (Fig. 3a) reveals that this CGG comprises two subdomains: one highly interconnected domain (CGG 2A) containing 38 genes that are loosely connected to a less interconnected group of 25 genes (CGG 2B). These two subdomains exhibit more coherent expression: CGG 2A in brain -0.63 ($P = 0.001$), kidney -0.79 ($P = 0.001$), and liver -0.74 ($P = 0.001$); and CGG 2B in brain -0.36 ($P = 0.013$), kidney 0.04 ($P = 0.15$), and liver 0.92 ($P = 0.001$). This illustrates the complexity of the correlations within the network where the existence of CGGs defined by correlation alone does not capture the full relationships of mRNA levels.

We again observed considerable intraindividual variation in the direction of CGG expression. None of the CGGs are coherently expressed in the same direction in all three tissues, except for CGG 2A (as described above), which is downregulated in all three tissues. The strength of coherency in each tissue coupled with the changing direction of expression again suggests that regulatory variation may be

Fig. 5 Coherency analysis. **a** Coherency overview: An example CGG containing 12 genes is identified by correlation analysis in the 31 BXD strains; the expression ratios from a comparison of two mouse strains for each of these 12 genes are shown (most genes are upregulated). The coherency score is calculated and statistical significance is determined via permutation (see Methods subsection “Coherency test statistic”). The resulting coherency and statistical significance are displayed as an annotated histogram. This process is repeated for all CGGs in expression data from all three tissues. **b** Intraindividual coherency: We plot the coherency scores for each CGG in the brain, kidney, and liver for DBA/2J vs. C57BL/6J in the first row (blue, green, and red, respectively) and for SJL/J vs. C57BL/6J in the second row (light blue, light green, and orange, respectively). **c** Interindividual coherency: The same data from panel B but reordered so that the tissues are grouped together. Stars indicate the degree of statistical significance (* $P < 0.05$, ** $P = 0.001$)



affecting both tissue-independent and tissue-specific factors, respectively, which combines to produce a tissue-specific outcome.

To confirm the previous observations in DBA/2J, we performed an independent comparison of a distinct inbred mouse strain, SJL/J, to C57BL/6J (see Methods subsection “RNA preparation: C57BL/6J vs. SJL/J experiment”). Because SJL/J is intermediate in nucleotide diversity between C57BL/6J and DBA/2J (Pletcher et al. 2004) and many inbred strains contain haplotypes from common ancestors (Frazer et al. 2007), we hypothesised that most but not necessarily all CGGs would behave like CGGs in this strain. We found that all CGGs are coherent in at least one tissue ($P < 0.05$; row 2 of Fig. 5b and Supplementary Table 1), and of the 12 possible cases of four CGGs measured in three tissues, four are highly coherent ($P < 0.001$) with four more with $P < 0.05$. Focusing on the tissues in which each CGG is coherent, we find a more consistent direction of expression of each CGG in this comparison of SJL/J to C57BL/6J than with DBA/2J vs. C57BL/6J. CGG 3 is coherently expressed in all three tissues ($P < 0.05$), CGGs 1 and 2 are coherently expressed in two tissues ($P < 0.001$), and there are no examples of CGGs that are coherent but expressed in different directions in two tissues (row 2 of Fig. 5b). These findings confirm that these groups of genes are indeed collectively sensitive to genetic influence, even in this more distant inbred strain. However, the greater similarity of the direction of CGG expression in multiple tissues of SJL/J compared to DBA/2J suggests that there are significant interindividual differences in the behaviour of CGGs.

To further investigate the interindividual coherency of CGGs, we compared the coherency and directionality of each CGG in DBA/2J and SJL/J, grouped by tissue in Fig. 5c. Of the seven cases in which a CGG is coherent in both DBA/2J and SJL/J, five are expressed in the same direction. The notable exceptions are CGG 3 in kidney and liver, which flip direction between DBA/2J and SJL/J, perhaps indicating that the causative regulatory variations were unique to the DBA/2J lineage, or that SJL/J contains additional regulatory variations that alter the expression levels of this CGG in kidney and liver. These findings demonstrate that CGGs tend to be more similarly expressed in the same tissue of different inbred mice than they are expressed in different tissues of the same inbred mouse.

CGG and biological functions

The existence of CGGs could be interpreted, at the extremes, as either the inevitable outcome of shared and partially shared mRNA level control or of a more specific regulatory architecture evolved to have functional outcomes. We therefore examined the genes in each CGG for

similarity of function. CGG 1 shows no obvious functional clustering. In contrast, CGG 2 contains a cluster of 12 genes (of 65) involved in metabolism, including six genes that encode proteins involved with complex carbohydrate metabolism. These genes include *Smc3*, a proteoglycan; *Muc2* and *Muc6*, heavily glycosylated proteins; *Acan*, a proteoglycan; *Chia*, involved in glycoside metabolism; and *Hyal2*, involved in glycosaminoglycan metabolism. Four additional genes encode proteins involved in sugar metabolism: *Gaa*, a glucosidase; *Glb1*, a galactosidase; *Akr1b3*, an aldose reductase; and *Slc2a8*, a facilitated glucose transporter. Two genes, *Aldh3a1* and *Aldh9a1*, encode proteins that are aldehyde dehydrogenase family members. Using enrichment analysis (see Methods subsection “Gene Ontology analysis”), CGG 2 is enriched for *tissue development* (GO:0009888; $P = 1.79 \times 10^{-3}$) and *carbohydrate metabolic process* (GO:0005975; $P = 1.83 \times 10^{-3}$), with the enrichments resulting from the gene sets *Acan*, *Spr2k*, *Gaa*, *Spr3*, *Bmp6* and *Acan*, *Chia*, *Gaa*, *Hyal2*, *Glb1*, *Slc2a8*, respectively. In contrast, CGG 3 contains five genes (of 21) that are involved in transcription control, including *Hoxc5*, *Zbtb43*, *Suv39h1*, *Foxq1*, and *Hes2* (but not supporting any statistically significant enrichment).

We found convincing evidence of functional clustering in CGG 2 and CGG 4. In CGG 4, 10 of the 14 transcripts are annotated [six are ribosomal proteins (*Rps29*, *Rps15*, *Rplp2*, *Rplp1*, *Rpl35A*, and *Rpl19*), and two are ribosomal protein/ubiquitin fusions (*Fau* and *Uba52*)] and showed a highly significant enrichment for *translation* (GO:0006412; $P = 2.77 \times 10^{-6}$).

It is plausible that patterning of coexpression that results in CGG composition could be simply secondary to genomic context, arising, for example, from coexpressed tandem repeat genes or gene pairs under common control due to the local chromatin environment (Batada et al. 2007; Fukuoka et al. 2004). To test this possibility, we examined the extent to which genes in are colocalised (defined as an adjacent pair of genes that are within 1 Mb of each other: see Methods subsection “Analysis of genomic location”). CGG 1 through CGG 5 had at least one gene on 20,18,13, 10 and 5 distinct chromosomes, respectively. Only 5 of 75 and 5 of 63 genes were spaced 1 Mb or less in CGG 1 and CGG 2, respectively. No such instances were observed on CGG 3, CGG 4, or CGG 5. On CGG 2, these colocalised gene pairs included *Spr2k* and *Spr3* and *Muc2* and *Muc6*, but no other pairs of genes from the same family were observed. Therefore, genomic localisation does not appear to be an organising principle of CGGs observed in our analysis.

We have stressed that the genetic influences upon CGGs do not have to be at the level of the control of transcription; nevertheless, this is a plausible hypothesis that is testable. To study this, we examined the CGGs for overrepresentation of transcription factor (TF) binding sites (TFBS); our

reasoning is that transcriptional control of a CGG could be due to shared action of TFs and that a variant TF could then contribute to the differential mRNA levels across our BXD panel. Our results are summarised in Supplementary Table 2 and here we discuss only CGG 2. We identified a number of enriched TFBS motifs that were present in more promoters of the 63 genes in CGG 2 than expected by chance ($P < 0.05$) (see Methods subsection “[Transcription factor binding motifs](#)”), suggesting that they may be involved in the regulation of the genes. These TFBS motifs correspond to 24 TFs, including *Foxd3*, *Tcf1*, *En1*, *Sp1*, *Gfi1*, *Nkx2-5*, *Irf2*, and all 17 TFs of the *Sox* family that recognise different variants of the AACAAAT motif (Koopman 2001) (*Sox1* to *Sox9*, *Sox11* to *Sox13*, *Sox15*, *Sox17*, *Sox18*, *Sox21*, and *Sox30*). If any of these TFs contribute to variation in CGG 2 mRNA levels, we may be able to detect genetic association of the TF gene with the mRNA levels of some or all of the genes in CGG 2. To identify association, we conducted a hypothesis-driven eQTL analysis in each of the three tissues to test for linkage of any of the 63 genes in CGG 2 to the closest genetic marker to each of the 24 TF genes identified above (see Methods subsection “[eQTL analysis for genes in CGG 2](#)”).

The marker *D8Mit124* located approximately 2.3 Mb distal of the *Sox1* gene on chromosome 8 had median a P value of 0.001 for the 63 mRNA levels in the brain compared to 0.410 for all other gene/TF marker combinations, 0.012 in the kidney compared to 0.422, and 0.015 in the liver compared to 0.488. While the individual P values do not reach significance following a Bonferroni correction, there is nevertheless a striking incidence of low P values to this marker. This result is compatible with the hypothesis that some of the variation in CGG 2 mRNA levels in all three tissues may be caused by genetic variation in the *Sox1* gene; however, the gene is located in a region of low polymorphism and there are no immediate candidate coding or noncoding SNPs. Proving involvement of *Sox1* will require an experimental design that is outside the scope of this study. Intriguingly, Blache et al. (2004) have previously suggested *Sox9* is a negative regulator of *Muc2* but did not test *Sox1*.

Finally, we tested the ability of CGG centroids to act as surrogate expression traits in genome-wide eQTL analyses, in a manner analogous to the PCA-derived *eigentrails* of Biswas et al. (2008). Overall, patterning linkage peaks across the genome was comparable between centroid- and per-gene analyses (data not shown). Only the CGG 2 centroid demonstrated linkage (in brain, $P = 0.044$ to *D1Mit134*, and in kidney, $P = 0.0011$ to 08.006.700). Interestingly, the latter marker was approximately 2.86 Mb proximal of the *Sox1* gene, suggesting that the centroid-based approach may have the ability to identify meaningful genetic determinants of expression *in trans*.

Discussion

In this study we have taken advantage of different genetic backgrounds to identify groups of genes whose mRNA levels are likely to be under shared genetic influences across multiple tissues. We focused on examining the inbred strains C57BL/6J and DBA/2J and limited our analyses of genetic influence to only those genes that were expressed in brain, kidney, and liver and that were differentially expressed between the parental strains in one or more of these tissues. We identified 755 genes subject to such genetic influence. Using pairwise comparisons of mRNA levels across 31 recombinant inbred strains of mice derived from this pair of parental strains, we detected “correlating groups of genes” or CGGs, whose mRNA levels change coordinately across all 31 strains in all three tissues. We then studied the same genes in the unrelated strain SJL/J and showed that they also exhibit CGG-like behaviour and exhibit coordinately up- or downregulated levels of mRNA, as appropriate.

Genetic influences that influence multiple genes must be formally *in trans*. We can therefore estimate the cumulative total influence on mRNA levels of *trans* effects as the proportion of the variation in an individual gene’s mRNA level that can be ascribed to shared CGG influences. Considering the average proportion of variation across a CGG, this quantity ranges from 52 to 79%, which is comparable in magnitude to reported results of eQTL analyses of effects which are *in cis* to a gene (Hubner et al. 2005; Petretto et al. 2006; Stranger et al. 2005; West et al. 2007). The influences we have observed on groups of genes are likely to be the outcome of numerous *in trans* influences that are individually not easy to detect. These data support the notion that *trans* influences, while being far less readily detectable than those *in cis*, are actually both common and of significant cumulative effect.

The organisation of genes into coordinately regulated groups or “modules,” based on both response to regulatory variation and broader disease or environmental challenge, has been recently reported by several groups (Chen et al. 2008; Keller et al. 2008). We see little overlap of genes in our CGGs and these authors’ modules, which have been defined by very different genetic backgrounds (data not shown). In our analysis, we further show that a striking feature of some CGGs is that genetic variation influences the same genes in divergent ways in different tissues of the same individual; genes in a CGG may be relatively upregulated in one (or more) tissue(s) but relatively downregulated in another. Unpredictable behaviour is also seen in the behaviour of CGGs compared across different individuals. For example, mRNAs of a CGG may be upregulated in the brain of one strain but downregulated in the brain of a second; we have observed this in replicated studies of

C57BL/6J, DBA/2J, and SJL/J, as well as in individual BXD strains. This unpredictability is quite unlike the effects of a protein sequence variation where an amino acid change is the same in every tissue that expresses the relevant exon.

We identify genetic influence in these studies by detecting pairs of genes whose mRNA levels vary coordinately in our analyses; however, the proportion of the 755 genes that are affected is entirely determined by the threshold used to construct the correlation network. Consistent with previous analyses (Brumm et al. 2008; Freeman et al. 2007), we have shown that there is no simple single criterion that we can use to define this cutoff (indeed there is no plausible biological reason why there should be a discrete value (Brumm et al. 2008; Wirth et al. 1966), but using the cutoffs employed for the three tissue analyses, we can show that approximately 80% of the 755 genes are genetically influenced in one or more tissues, suggesting that these complex *trans*-genetic influences are common. It is also likely that there are groups of coregulated genes that would not have been included in our initial 755-gene analysis but that are revealed as genetically influenced because they are subject to transgressive segregation in the BXD strains. We highlight that there are many different approaches that, in principle, have been used to construct an expression correlation network, e.g., the weighted approach of Horvath and colleagues (Emilsson et al. 2008; Zhang and Horvath 2005), Bayesian networks (Bansal et al. 2007), or information theory-based approaches (Bansal et al. 2007; Reverter and Chan 2008). More importantly, because no method of network analysis is without its limitations or disadvantages, we pursued coherency analysis as a method for assessing our analysis in the context of independent experiments.

The apparently common but unpredictable influence of genetic variation prompted us to develop the use of coherency testing, essentially testing the direction rather than the amount of relative change in mRNA levels for analysis of relative CGG gene behaviour. We believe this is a robust and appropriate test of a CGG that is not based on the extreme view that mRNA levels should be identical between two genetically dissimilar individuals. Further extensions to the present methods of coherency testing are also possible. Our current approach is limited to testing the extent to which groups of genes show uniform changes in expression, but if more complex patterns of coregulation could be specified, these approaches could remain informative.

Our data add to three lines of evidence suggesting that the influence of genetic variation is frequently tissue-specific. First, several microarray-based surveys have highlighted differences in gene expression across different brain regions in inbred mouse strains (Freeman et al. 2007; Hovatta et al. 2007; Nadler et al. 2006; Pavlidis and Noble 2001; Sandberg et al. 2000). These differences in

expression appear to be phenotypically relevant, as shown by analysis of interstrain differences in motor coordination tasks (Nadler et al. 2006). Second, analyses of eQTL data from studies on different tissues have shown limited evidence for tissue-specific effects (Bystrykh et al. 2005; Chesler et al. 2005; Gatti et al. 2007; Hubner et al. 2005; Petretto et al. 2006). Third, Yang et al. (2006), using an intercross of C57BL/6J and C3H/HeJ mouse strains and sampling muscle, liver, adipose, and brain, demonstrated the essentially tissue-specific nature of expression of sexually dimorphic, but not more general, classes of genes.

Functional annotation of genes within each CGG showed that in some cases genes whose mRNA levels were highly correlated also encode proteins with biologically related functions; the clearest examples are proteins involved in sugar metabolism clustered in CGG 2 and ribosomal proteins in CGG 4. The correlated behaviour of functionally related genes is perhaps not surprising in view of numerous studies on the coregulation of gene expression. Our major conclusion, however, is that shared function does not appear to be the primary organising principle of most genes within a CGG. In this respect, a better understanding of the shared behaviour of the CGG and its relationship, if any, to phenotypic outcomes (Goring et al. 2007; Nadler et al. 2006; Passador-Gurgel et al. 2007) will provide greater insight into the functional consequences of CGG variation and shared control. Further insight into the putative biology of CGGs could be gained by employing them as “gene-sets” in pathway-based enrichment analyses, such as Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005), across a variety of experiments.

Our findings for CGG 2 that the *Sox* binding site is overrepresented and that mRNA levels of the genes within the CGG exhibit unusual linkage at the region harbouring *Sox1* suggest an involvement of this transcription factor in CGG 2 behaviour, but this is necessarily speculative. The reality is that our methods, in common with all such analyses, including eQTL-based approaches, cannot distinguish between primary and secondary influences upon mRNA levels. For example, whether an unobserved common regulator causes CGG 2 behaviour or variation in more distal processes, such as signal transduction, will have to be shown by extensive mechanistic dissection, but such follow-up studies will at least have to be able to distinguish between these alternatives.

In more general terms, we have focused upon correlation-based approaches in our study with the assumption that correlation is a likely outcome of biological processes rather than simply using correlation as a statistical tool. This study has not been designed to identify, in most cases, the cause of a change in mRNA level; instead, we have simply focused on defining at the level of mRNA the phenotypic differences between two organisms that are

likely due to the sum total of all relevant genetic influences. Of course, changes in mRNA levels do not have to be reflected in changing protein levels, and in most cases it is this latter change that will contribute to phenotypic diversity. Recent studies in yeast by Foss et al. (2007) have shown that there is only weak correlation of mRNA and protein levels tested across genetically divergent strains, and so prediction from purely genotypic information of ultimate protein levels, and therefore potential phenotype, is going to be a very challenging task, even at a single-tissue level alone at a multiple-tissue or organismal level. Nevertheless, the observation that this type of genetic variation has strong tissue-specific outcomes suggests that the regulatory architecture of mRNA levels may have evolved, in part, to generate selective phenotypic diversity of individual tissues and could represent a contributing source of morphologic and functional evolutionary differences, as well as having implications for the use of surrogate tissues in human studies.

Acknowledgments We thank Maja Bucan and John Schimenti for assistance with obtaining the BXD mice. We also thank Grant Morahan for access to BXD strains, Ruby Lin for technical advice on microarrays and RNA extraction, and Peter Koopman for advice regarding the *Sox* family of TFs. MJC and EKFC were supported by Australian Postgraduate Awards; CJC was supported by a School of Biotechnology and Biomolecular Sciences Genome Information Scholarship; RBHW was supported by a National Health and Medical Research Council of Australia Peter Doherty Fellowship; and the project work was funded by an Australian Research Council grant to PFRL. High-performance computing resources were provided by the Australian Partnership for Advanced Computing (APAC) and the Australian Centre for Advanced Computing and Communications (ac3). We thank Stylianos Antonarakis and Rick Myers for insightful comments on earlier drafts of the manuscript, and the two anonymous reviewers whose comments greatly improved our manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D (2007) How to infer gene networks from expression profiles. *Mol Syst Biol* 3:78
- Batada NN, Urrutia AO, Hurst LD (2007) Chromatin remodelling is a major source of coexpression of linked genes in yeast. *Trends Genet* 23(10):480–484
- Biswas S, Storey JD, Akey JM (2008) Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 9:244
- Blache P, van de Wetering M, Duluc I, Domon C, Berta P et al (2004) SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the *CDX2* and *MUC2* genes. *J Cell Biol* 166(1):37–47
- Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* 102(5):1572–1577
- Brumm J, Conibear E, Wasserman WW, Bryan J (2008) Discovery and expansion of gene modules by seeking isolated groups in a random graph process. *PLoS ONE* 3(10):e3358
- Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* 41:8–429
- Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT et al (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nat Genet* 37(3):225–232
- Chen Y, Zhu L, Lum PY, Yang X, Pinto S et al (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature* 452(7186):429–435
- Chesler EJ, Lu L, Wang J, Williams RW, Manly KF (2004) WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat Neurosci* 7(5):485–486
- Chesler EJ, Lu L, Shou S, Qu Y, Gu J et al (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 37(3):233–242
- Cotsapas CJ, Williams RBH, Pulvers JN, Nott DJ, Chan EK et al (2006) Genetic dissection of gene regulation in multiple mouse tissues. *Mamm Genome* 17(6):490–495
- Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F et al (2008) Genetics of gene expression and its effect on disease. *Nature* 452(7186):423–428
- Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A et al (2007) Genetic basis of proteome variation in yeast. *Nat Genet* 39(11):1369–1375
- Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA et al (2007) A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* 448(7157):1050–1053
- Freeman TC, Goldovsky L, Brosch M, van Dongen S, Mazière P et al (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol* 3(10):2032–2042
- Frith MC, Fu Y, Yu L, Chen JF, Hansen U et al (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32(4):1372–1381
- Fruchterman TM, Reingold EM (1991) Graph drawing by force-directed placement. *Softw Exper Pract* 21(11):1129–1164
- Fukuoka Y, Inaoka H, Kohane IS (2004) Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. *BMC Genomics* 5(1):4
- Gatti D, Maki A, Chesler EJ, Kirova R, Kosyk O et al (2007) Genome-level analysis of genetic regulation of liver gene expression networks. *Hepatology* 46(2):548–557
- Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5(10):R80
- Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C et al (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* 2(8):e130
- Gibson G, Weir B (2005) The quantitative genetics of transcription. *Trends Genet* 21(11):616–623
- Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J et al (2007) Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet* 39(10):1208–1216
- Gower JC, Ross GJ (1969) Minimum spanning trees and single linkage cluster analysis. *Appl Stat* 18:54–64
- Hovatta I, Zapala MA, Broide RS, Schadt EE, Libiger O et al (2007) DNA variation and brain region-specific expression profiles

- exhibit different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biol* 8(2):R25
- Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H et al (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* 37(3):243–253
- Ihaka R, Gentleman RC (1996) R: a language for data analysis and graphics. *J Comput Graph Stat* 5:299–314
- Keller MP, Choi Y, Wang P, Davis DB, Rabaglia ME et al (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res* 18(5):706–716
- Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G et al (2007) Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci USA* 104(5):1708–1713
- Koopman P (2001) SRY and DNA-bending proteins. *Encyclopedia Life Sci* 1–6
- Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ et al (2007) The UCSC genome browser database Update. *Nucleic Acids Res* 35(Database issue):D668–D673
- Lönnstedt I, Speed TP (2002) Replicated microarray data. *Stat Sin* 12:31–46
- Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA et al (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat Genet* 37(11):1224–1233
- Nadler JJ, Zou F, Huang H, Moy SS, Lauder J et al (2006) Large-scale gene expression differences across brain regions and inbred strains correlate with a behavioral phenotype. *Genetics* 174(3):1229–1236
- Passador-Gurgel G, Hsieh WP, Hunt P, Deighton N, Gibson G et al (2007) Quantitative trait transcripts for nicotine resistance in *Drosophila melanogaster*. *Nat Genet* 39(2):264–268
- Pavlidis P, Noble WS (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol* 2(10):RESEARCH0042
- Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK et al (2006) Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet* 2(10):e172
- Pletcher MT, McClurg P, Batalov S, Su AI, Barnes SW et al (2004) Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse. *PLoS Biol* 2(12):e393
- Reverter A, Chan EK (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24(21):2491–2497
- Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nat Rev Genet* 7(11):862–872
- Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA et al (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc Natl Acad Sci USA* 97(20):11038–11043
- Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3
- Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S et al (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet* 1(6):e78
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102(43):15545–15550
- Taylor BA, Wnek C, Kotlus BS, Roemer N, MacTaggart T et al (1999) Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. *Mamm Genome* 10:335–348
- Vlieghe D, Sandelin A, De Bleser PJ, Vlemincx K, Wasserman WW et al (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res* 34(Database issue):D95–D97
- Voy BH, Scharff JA, Perkins AD, Saxton AM, Borate B et al (2006) Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol* 2(7):e89
- West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelsmore RW et al (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175(3):1441–1450
- Williams RB, Chan EK, Cowley MJ, Little PF (2007) The influence of genetic variation on gene expression. *Genome Res* 17(12):1707–1716
- Wirth M, Estabrook GF et al (1966) A graph theory model for systematic biology, with an example for the *Oncidiinae* (Orchidaceae). *Syst Zool* 15:59–69
- Wu C, Delano DL, Mitro N, Su SV, Janes J et al (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet* 4(5):e1000070
- Yang YH, Dudoit S, Luu P, Speed TP (2001) Normalization for cDNA microarray data. In: Bittner ML, Chen Y, Dorsel AN, Dougherty ER (eds) *Microarrays: optical technologies and informatics*, vol 4266 of *Proceedings of SPIE*, pp 141–152
- Yang X, Schadt EE, Wang S, Wang H, Arnold AP et al (2006) Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res* 16(8):995–1004
- Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Stat Appl Genetics Mol Biol* 4: Article17