Intra– and inter– individual genetic differences in gene expression

Mark J. Cowley[1,4], Chris J. Cotsapas[1,2,4], Rohan B. H. Williams[1,2,4],
Eva K. F. Chan[1], Jeremy N. Pulvers[1], Michael Y. Liu[1], Oscar J. Luo[1,2]
David J. Nott[3] and Peter F.R. Little[1,2]
[1]School of Biotechnology and Biomolecular Sciences
[2]The Clive and Vera Ramaciotti Centre for Gene Function Analysis
[3]School of Mathematics and Statistics
The University of New South Wales Sydney, NSW AUSTRALIA

[4]Thsee authors contributed equally to this work
Corresponding author: peter.little@nus.edu.sg

Abstract

Genetic variation is known to influence the amount of mRNA produced by a gene. Given that the molecular machines control mRNA levels of multiple genes, we expect genetic variation in the components of these machines would influence multiple genes in a similar fashion. In this study we show that this assumption is correct by using correlation of mRNA levels measured independently in the brain, kidney or liver of multiple, genetically typed, mice strains to detect shared genetic influences. These correlating groups of genes (CGG) have collective properties that account for 40–90% of the variability of their constituent genes and in some cases, but not all, contain genes encoding functionally related proteins. Critically, we show that the genetic influences are essentially tissue specific and consequently the same genetic variations in the one animal may up–regulate a CGG in one tissue but down–regulate the same CGG in a second tissue. We further show similarly paradoxical behaviour of CGGs within the same tissues of different individuals. The implication of this study is that this class of genetic variation can result in complex inter– and intra–individual and tissue differences and that this will create substantial challenges to the investigation of phenotypic outcomes, particularly in humans where multiple tissues are not readily available.

Introduction

Gene expression is controlled by multiple molecular machines whose interaction with a gene and genes transcript contributes to determining a final level of mRNA; recent studies have shown that these processes are subject to significant influences of genetic variation that result in heritable changes to final mRNA levels (reviewed by Cotsapas *et al.*, 2006; Gibson and Weir, 2005; Rockman and Kruglyak, 2006; Williams *et al.*, 2007). In multicellular organisms, these molecular machines are involved in setting mRNA levels of many genes but they contain different components; some may be common to the expression of all genes in all cells of the organism whereas other components may have more limited function such that that are involved with sub sets of genes or sub sets of cell types, or both (Maniatis and Reed, 2002; Tsankov *et al.*, 2006; Maciag *et al.*, 2006; Komili and Silver, 2008). We therefore predict that genetic perturbation of these machines will result in either global or cell–type specific changes to gene expression, depending on the variant component. Such behaviour is in marked contrast to genetic variation in protein coding sequence, where the variant is observed in all cases where the gene is expressed. In this work, we use correlation based methods to show that the effects of regulatory variation are, as predicted, coordinated changes to the mRNA levels of groups of genes. These group changes can be very different both in multiple tissues of the same individual, as well as being different in the same tissues of multiple individuals. We use the term regulatory variation 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 through changes to *trans*–acting components, including any of the molecular machinery that controls the amount of steady state mRNA in a cell, such as transcription or splicing factors (Williams *et al.*, 2007). The majority of findings to date, using predominantly expression QTL (eQTL) experimental designs, suggest that *cis*–acting regulatory variation appears to be of larger effect size, and is thus more easily detected; in comparison, *trans*–regulatory variation appears to be of smaller effect size, and are either less common, or harder to detect (Petretto *et al.*, 2006; Stranger *et al.*, 2005; Goring *et al.*, 2007). When *trans*–acting influences are identified, these tend to be a small number of eQTLs that influence the expression of large numbers of genes, so called "master–regulator" of gene expression, suggesting that regulatory variation is affecting the expression level of groups of genes, simultaneously. Investigating *trans*–acting regulatory variation using eQTL analysis is presently beset both by very substantial statistical problems of multiple hypothesis testing and by the sheer scale of studies required to provide genetic power to detect small effect sizes. Further, whilst eQTL analysis is an appropriate approach to investigate the effects caused by one or a small number of genetic influences, it has limited power to detect additional eQTLs with smaller effect sizes (Brem and Kruglyak, 2005; Williams *et al.*, 2007). To 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 (Ghazalpour *et al.*, 2006; Lan *et al.*, 2006). 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 have the added advantage that we are measuring the shared outcome of regulatory variation stemming from multiple genetic loci, the modest contributions of which eQTL analysis would be underpowered to detect in all but the largest studies.

In this study of inbred and recombinant inbred mice, we set out to investigate *trans*–acting regulatory variation, using correlation analysis to identify groups of genes that are likely to be influenced by shared regulatory variation, and thus shared regulatory factors. We further investigate the consequence of *trans*–acting regulatory variation in three different mouse tissues to assess the degree to which *(A)* genes are affected by the same regulatory variation in all tissues, and *(B)* whether the outcome of such regulatory variation is the same in all tissues.

Overview of experimental design

To identify genes whose expression levels may be affected by regulatory variation, and to investigate their regulation in multiple tissues, we adopt the following experimental design: first, we compare gene expression levels in 3 tissues of two inbred mouse strains, C57BL/6J and DBA/2J, and 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 co–ordinately across the BXD RI strains and the three tissues; we term these "correlating groups of genes" or CGGs. We 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. We further investigate the specific outcomes of the regulation of these CGGs in the three tissues of the BXD panel.

Identify genetically influenced genes

We began by identifying genes differentially expressed in at least one of whole brain, kidney, or liver between strains C57BL/6J and DBA/2J. We found that we could reliably detect 6075 transcripts above background in all three tissues, of which 755 were variantly expressed between the two strains at a LOD>3 (the B–statistic of Lonnstedt and Speed (2002), as modified by Smyth (2004); see **Methods**). 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 selecting genes that are expressed in a "tissue specific" manner, in the sense of being expressed in only 1 or 2 of the 3 tissues (**Supplementary Figure 1**).

The identification of 755 genes as being potential targets of regulatory variation(s) does not us allow us to identify 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 highly correlated alterations mRNA levels of otherwise unrelated genes. 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 ask if the outcome of genetic influence on genes is the same in each tissue.

Identifying groups of genes with similar expression patterns in multiple tissues

To achieve this, we measured mRNA levels of the 755 genes in the same 3 tissues in 31 BXD recombinant inbred (RI) strains (Taylor *et al.*, 1999), pooling three age– and sex–matched mice from each. These strains have been derived from crosses of C57BL/6J and DBA/2J, which have been bred to homozygosity by repeated sibling pair mating. As they carry arbitrary mixtures of the two progenitor backgrounds, but are homozygous at each locus, we predict 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.

In order to identify those genes that have similar expression patterns in all 31 BXD strains and in all 3 tissues, we adopted a correlation–network approach (see **Methods**). We compare all pairwise combinations of 755 gene expression patterns across the 93 measurements, and construct correlation networks consisting of nodes representing genes, and edges representing correlations that are stronger than an empirically determined threshold.

The correlation–network approach has a number of advantages: the resulting networks summarise a large amount of complex data in a form that is easily visualised and interpreted, and there are a number of techniques for identifying discrete regions of the network corresponding to CGGs. Most importantly, the number of groups does not need to be known a priori as in clustering methods, and those genes that are not correlated highly enough are automatically filtered from the resulting network, thus reducing the noise in the system.

Choice of threshold for network construction and network properties

To display the complex relationship between genes and multiple pairwise correlations we construct networks using the widely used approach of thresholding correlation matrices (Freeman *et al.*, 2007; Voy *et al.*, 2006). The intention of thresholding is to define discrete groups of genes that can be subject to other analyses but we stress that there is no plausible reason why a threshold should have an explicit biological meaning, as regulation–induced correlation can be of any magnitude. An important step in constructing such networks is choice of threshold: too low a threshold will result in a too densely connected graph, while too high a threshold will result in a sparsely populated and connected network (Freeman *et al.*, 2007). The final choice of threshold is guided entirely by the overall objectives of the analysis. Our primary aim is to identify groups of co–regulated genes that are plausibly under common genetic control, so we focus on finding groups of interconnected genes that are distinct from other such groups (connected components in graph theoretical terms, unconnected to others).

We focused on identifying a correlation threshold that would *(A)* provide an adequate number of connected components that had at least 2 genes; *(B)* distinguish the graph–theoretic properties of the 755 genes from those of all expressed genes. To do so, we studied various network properties of these genes, treating them as test–statistics, and examined how unusual these properties were using 1000 sets of the same size randomly sampled from the 6075 genes that were expressed in all three tissues. We studied 6 graph–theoretic properties in this fashion (**Figure 1**), namely: *(1)* the number of correlations above the threshold; *(2)* the number of connected components; *(3)* the median of the distribution of connected component size; *(4)* size of the largest connected component, *(5)* average degree, computed across non–singleton connected components and *(6)* the global clustering coefficient (for a review of these concepts Sharan *et al.*, 2007).

At low correlation thresholds, we found that both the set of 755 genes ("755 network") and randomly resampled sets of genes ("random networks") formed networks characterised by a single, large connected component. This expected structure starts to break down into multiple connected components at $|\rho| > 0.50$ in the 755–network, and at $|\rho| > 0.35$ in random networks, with the former having consistently higher number of connected genes (**Figure 1A**). The number of connected components was also on average higher in the 755 network than in the random networks, but only at $|\rho| = 0.85$ was this greater than for all random networks. The largest number of connected components for both the 755– and random networks was observed at a threshold of $|\rho| = 0.75$, after which some of these structures become completely unconnected and disappear (**Figure 1B**). The median number of genes in connected components showed little difference between the 755–network and random ones (**Figure 1C**), but the size of the largest connected component in the 755 network was consistently above that observed for random networks (**Figure 1D**) across a wide threshold range. The average number of connections of genes in any component (their degree) of the 755–network was consistently higher than for random networks (**Figure 1E**), suggesting tighter overall correlation. The extent of connectivity within the component to which a gene belongs (measured by the clustering coefficient) was also higher for the 755 network at $0.55 \leq |\rho| \leq 0.75$; however, this measure becomes erratic above 0.775 due to the reduced size of the network at these stringent thresholds (**Figure 1F**).

In summary, the observed 755–network generated consistently higher number of correlations across a range of thresholds, generated a higher average level of connectivity between genes and a greater level of inter–connectivity between the neighbours of a given gene, than was observable in random networks of the same size derived from all expressed genes, suggesting that these genes are indeed responding to the influences of regulatory variation. We settled on a threshold of $|\rho| = 0.775$, which gave us maximal differences between the 755–network and background without dissolving structure due to high stringency.

The cross tissue correlation network

We constructed a $|\rho| = 0.775$ correlation network containing 212 (28.1%) genes that correlate with at least one other transcript; the genes have a median degree of 4, with 73% of genes with a degree of ≥ 2 (**Figure 2**). These genes are central to our subsequent study; in principle they are influenced by genetic 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 a further 204 (27.0%) genes are correlated in any pair of tissues, and a further 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).

In the original network across all tissues, we find that the 212 genes fall into 19 discrete correlating groups of genes or CGGs; of these groups, 10 contain at least three members and the largest 5 contain 75, 63, 21, 12 and 6 genes respectively; all 19 CGGs are displayed in **Figure 2** along with their expression patterns across the 31 BXD lines and the 3 tissues.

Given that CGGs are constructed from combinations of pair–wise relationships between genes, we expected that the levels of each transcript within a CGG (grey lines in **Figure 2**) should in general be similar. To assess the extent to which variation in a single genes mRNA could be explained by the shared influences upon a CGG, we correlated (using the coefficient of determination, R^2, see **Methods**) the expression pattern of each transcript to the centroid of their respective CGG (thick coloured lines in **Figure 2**) for each tissue individually, and across all 3 tissues simultaneously; we determined the statistical significance of the observed R^2 via permutation (see **Methods**). R^2 values ranged from ~ 0.4 to ~ 0.9 (**Figure 3**) and in most cases there was very limited overlap with randomly sampled genes. The permutation analysis shows that the behaviour of the genes in the CGG cannot easily be explained by inter-array differences in hybridisation; in this case we would expect the random permutations also to generate higher R^2 values. We note that even though CGG 1 in the brain has the smallest R^2 values, which overlap with randomly sampled genes, this CGG nevertheless exhibits significantly unusual biological behaviour that independently supports the notion it is a CGG (see below). We conclude from this analysis that shared behaviour is a significant influence upon genes within CGGs and that the shared influence upon genes in CGG (∼40–90% from **Figure 3**) is comparable in magnitude to the size of effects reported for *cis*–acting eQTLs Hubner *et al.* (2005); Petretto *et al.* (2006); Stranger *et al.* (2005); West *et al.* (2007).

While we have illustrated the congruous behaviour of mRNAs within a CGG, we also note from **Figure 2** that mRNA level profiles between each tissue are strikingly different. This is supported by calculating the correlation between the intra–tissue centroids for each CGG (**Table 1**): the only statistically significant relationship is in fact an anti–correlation between the centroids of CGG 2 in Brain and Liver ($\rho = -0.59$, $P = 5.94 \times 10^{-4}$). These results show that whilst 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 anti-correlated. These differences are best explained by genetic variation in multiple regulatory components that act individually in a tissue specific fashion or in a single cross tissue component whose behaviour is itself modulated by tissue specific factors.

The collective behaviour of CGGs

We have identified CGGs based on their expression patterns across a panel of BXD mice, and across three tissues. 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 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 identical expression of all genes in the CGG, we designed a test for the identical direction of mRNA levels: relatively up– or down–regulated. This co-ordinated expression over all genes in a CGG can be summarised as a coherency score: the proportion of genes whose mRNA levels are up–regulated relative to the reference (see **Figure 4a** for an example, and **Methods** for details).

We performed simulation studies to assess the performance of the coherency score with respect to both the number of genes in a CGG, and the magnitude and variability of the expression changes (see **Supplementary Material**). 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 10 genes (at permuted $P < 0.05$). Below this group size, the score had little power even in the case of maximal coherency.

We applied this method to the 4 largest CGGs (those having between 75 and 12 genes). Given the CGG had been defined solely by analysis of the BXD RI strains, we therefore looked at coherency in the two progenitor strains, C57BL/6J and DBA/2J and found that all 4 CGGs were significantly coherent ($P = 0.001$) in at least one tissue (**Figure 4b** and **Supplementary Table 1**). 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, $P = 0.001$); whilst the shared contribution to overall mRNA levels of the CGG might be relatively small, there is a marked effect upon direction of mRNA level changes. We also note that the 63 genes in CGG 2 have complex properties: coherency is moderate in size, 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 (**Figure 2**) reveals that this CGG comprises two sub domains, one highly interconnected domain (CGG2A) containing 38 genes, which are loosely connected to a less interconnected group of 25 genes (CGG2B). These two sub-domains exhibit more coherent expression: CGG2A in Brain -0.63 ($P = 0.001$), Kidney -0.79 ($P = 0.001$) and Liver -0.74 ($P = 0.001$) and CGG2B in Brain -0.36 ($P = 0.013$), in 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.

To validate these observations, we performed an independent comparison of a distinct inbred mouse strain, SJL/J to C57BL/6J (see **Methods**). Given the close genetic relationship between DBA/2J and SJL/J (Beck *et al.*, 2000), we expected that these 4 CGGs should *(A)* be coherent in each tissue, and *(B)* show similar directionality as DBA/2J with respect to C57BL/6J. We found both predictions to be true, with all CGGs being coherent in at least one tissue ($P < 0.05$; see **Figure4b, row 2**, and **Supplementary Table 2**), and most CGGs that were coherent in both DBA/2J and SJL/J having the same directionality in the same tissue (those entries with ** for both strains in **Figure 4c**). These findings confirm that these groups of genes are indeed collectively sensitive to genetic influence, even in this more distant inbred strain.

Inter–strain and inter–individual coherency

These observations provide independent biological confirmation of the properties of CGGs, and also reveal the complex outcomes of genetic influence upon mRNA levels. In these 3 inbred strains, mRNA levels of the same groups of genes co–ordinately vary not just between strains but also between the same tissues between each strain. If this behaviour is indeed genetic in origin, as we have argued, then we would also expect the same to be true in the BXD lines. We therefore calculated coherency for the four largest CGGs in the three tissues of each BXD line relative to C57L/6J and compared this to the coherency of DBA/2J expression. Due to the small number of replicate microarrays used for each measure (6 for DBA/2J, 3 for SJL/J and 1 for each BXD RI), we limit our inferences of differential regulation to extreme cases, where a high score in DBA/2J becomes of high magnitude but opposite sign in at least one BXD strain. We note that a simple Mendelian effect would result in either a coherency score approximating $+/-1$ for a DBA/2J allele or 0 for a C57BL/6J allele, which is the reference strain. We use permutation to assign significance to these events, assessing the likelihood that a given coherency would occur by chance in our dataset (see **Methods**).

We find CGG 2 in kidney changes from -0.46 in DBA/2J to $+1.00$ in 2 BXD lines; CGG3 in brain changes from $+0.33$ to -1.00 in 3 BXD lines, in kidney from $+1.00$ to -1.00 in an BXD line, in liver from -1.00 to $+1.00$ in 7 BXD lines; CGG 4 in brain from $+0.83$ to -1.00 in 4 BXD lines, in kidney from -0.83 to $+1.00$ in 6 BXD lines. The pattern of changes in coherency of CGG 3 in kidney shown in **Figure 5A** is compatible with segregation of variations within the BXD lines and the consistent up–regulation of mRNA levels in the kidney for most of the BXD lines for CGG 4 (**Figure 5B**) is supportive of transgressive segregation of genetic influences. (see **Supplementary Table 3** for full coherency scores from the BXD panel, DBA/2J and SJL/2J). Collectively, these results suggest that genetic variation has influences that result in the effective tissue specificity of changes in mRNA levels and that even the direction of this change is not readily predictable either within or between individuals.

This analysis identifies dramatic changes in coherency of CGGs across the panel, supporting the genetic origin of this phenomenon, and allows us to define some extreme coherency alterations that are likely segregating within the BXD lines. However, we acknowledge the lack of power to draw more specific conclusions as to the full range of coherency phenotypes displayed across the complete strain collection.

Encoded protein functions and CGG identity

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. To address this latter possibility, we sought to find functional relationships within CGGs, whose sizes allow for statistically valid analyses, using Gene Ontology (GO) Biological Process terms (Ashburner *et al.*, 2000). We found convincing evidence of functional clustering in CGG 2 and CGG 4. In CGG 4, ten of the fourteen 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; $P = 2 \times 10^{-6}$). CGG 2 contains 63 genes, 35 with GO annotations: 13 are involved in carbohydrate metabolism, 5 involved in signalling and 4 involved in transport and was enriched for *carbohydrate metabolic process* (GO:0005975; $P = 2.1 \times 10^{-4}$). We note that CGG 2 illustrates the complexity of breaking a network into discrete sub–networks: whilst we can analyse CGG 2 as a whole, there remains distinct functional clustering even within the CGG. These findings are compatible with some CGGs having functional significance but certainly do not support the view that shared function is the major determinant of CGG gene content.

We have stressed that the genetic influences upon CGGs do not have to be at the level of the control of transcription but this is nevertheless a plausible hypothesis that is testable. To study this, we examined the CGGs for over-representation 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 4** and here we discuss only CGG 2; we identified 24 TFs, including *FOXD3*, *TCF1*, *EN1*, *SP1*, *GFI1*, *NKX2–5*, *IRF2*, and 17 TFs of the *Sox* family (*Sox1* to *Sox9*, *Sox11* to *Sox13* and *Sox15*, *Sox17*, *Sox18*, *Sox21* and *Sox30*) whose cognate binding sites were present in more of the promoters of the 63 genes in CGG 2 than expected by chance ($P < 0.05$) (see **Methods**) suggesting they may be involved in the regulation of the genes. If any of the TFs are contributing 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 carried out an eQTL analysis across all 3 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**).

The marker D8Mit124 located ∼2.3Mb distal of the Sox1 gene on chromosome 8 had median $P-$values 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. Whilst the individual $P-$values do not reach significance under 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 or protein: the gene is located in a region of low polymorphism and there are no immediate candidate coding or non-coding SNPs. Proving involvement of *Sox1* will require an experimental design that is outside the scope of this study.

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 have focused on examining the inbred strains C57BL/6J and DBA/2J and limited our analyses of genetic influence only to those genes that were expressed over a defined mRNA level in brain, kidney and liver and that were differentially expressed between the parental strains in one or more of these tissues: we identify 755 genes subject to such genetic influence. Using pairwise comparisons of mRNA levels across 31 recombinant inbred lines of mice derived from this pair of parental strains, we detect "correlating groups of genes" or CGGs, whose mRNA levels change co–ordinately across all 31 strain 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 co–ordinately up– or down–regulated levels of mRNA, as appropriate. We further show a striking feature of some CGGs is that genetic variation influences the same genes in divergent fashions in different tissues of the same individual; genes in a CGG may be relatively up–regulated in one (or more) tissue(s) but relatively down–regulated in another. Unpredictable behaviour is also seen in the behaviour of CGGs compared across different individuals: for example, mRNAs of a CGG may be up–regulated in the brain of one strain but down–regulated in the brain of a second and 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 co–ordinately in our analyses; however, the proportion of the 755 genes that are affected is entirely determined by the cut–off used to construct the correlation network. Consistent with previous analyses (Freeman *et al.*, 2007), we have shown that there is no simple single criterion that we can use to define this cut–off (indeed there is no plausible biological reason why there should be a discrete value) but using the cut–offs employed for the three tissue analyses, we can show that 80% of the 755 genes are genetically influenced in one or more tissues, suggesting these complex genetic influences are common. It is also likely there are groups of co–regulated genes that would not have been included in our initial 755– gene analysis but that are revealed as being genetically influenced due to their being subject to transgressive segregation in the BXD lines. The apparently common but unpredictable influence of genetic variation prompted us to develop the use of coherency testing, essentially testing the direction rather than amount of relative mRNA levels change, for analysis of relative CGG gene behaviour. We believe this is a robust and appropriate test of a CGG that is not based upon the extreme view that mRNA levels should be identical between 2 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 co–regulation could be specified, these approaches could remain informative.

Our data adds to three lines of evidence suggesting that the influence of genetic variation is frequently tissue specific. Firstly, 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) and these differences in expression appear to be phenotypically relevant, as shown by analysis of inter–strain differences in motor coordination tasks (Nadler *et al.*, 2006). Secondly, analyses of eQTL data from studies in 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). Thirdly, Yang *et al.* (2006), using an inter–cross 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 13 proteins involved in sugar metabolism clustered in CGG2 and 6 ribosomal proteins in CGG4. The correlated behaviour of functionally related genes is perhaps not surprising in view of numerous studies on the co–regulation 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.

The proportion of the variation in an individual genes mRNA level that can be ascribed to shared CGG influences ranges from 40-90%, which is very similar to reported results of eQTL analyses, in particular 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). Logically, influences shared between 2 or more genes are difficult to reconcile with *cis* acting variations and the smaller effects on mRNA levels of the *trans* acting influences detectable in eQTL studies suggests that the correlation influences we detect are the outcome of numerous, additive, *in trans* influences that are individually not easy to detect. We note that our study design is, like most other published accounts, underpowered to detect significant eQTLs at a whole genome scale and we have therefore not attempted this approach at a global level. We do however provide evidence that multi-factorial trans–acting genetic variants must exist; appearing to influence gene groups of modest size, as supported by previously published eQTL analyses.

Steady–state mRNA levels are set by a complex set of regulatory interactions, only some of which will be primary modulations of transcription. Our findings for CGG 2 that the *Sox* binding site is over-represented and 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 minimally 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 but rather we have simply focused upon 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 from Foss *et al.* (2007) have shown 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, let 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 morphological and functional evolutionary differences.

Finally, if tissue specificity of genetic influence is replicated in humans, then using mRNA levels measured in readily available surrogate tissues will not easily predict outcomes in more relevant tissues and this will have very substantial implications for the design of human studies.

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, 27–29, 30–34, 36, 38, 39, 40, and 42 were obtained from the Jackson Laboratories (Bar Harbor, ME, USA). 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 N_2. Total RNA was extracted according to the manufacturer's instructions with TRIzol Reagent (Invitrogen, Mt. Waverley, Vic, Australia); purity and integrity was assessed by OD_{260}/OD_{280} readings greater than 2 and intact rRNA bands (Agilent Bioanalyzer, Agilent, Forest Hills, Vic, Australia) analysis, respectively.

Parental strain experiment: Total RNA from the three tissues of 10 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, Mt. Waverley, Vic, Australia) and fluorescently labelled with Alexa Fluor 555 for C57BL/6J and Alexa Fluor 647 for DBA/2J (Invitrogen, Mt. Waverley, Vic, Australia).

BXD panel experiments: Equal amounts of total RNA from 3 animals from each BXD strain were mixed to give tissue pools representative of the genetic backgrounds. A common reference sample was created for each tissue from total RNA extracted from ten eight–week–old male C57BL/6J mice (a different RNA source than the parental strain experiment). 20 μg of pooled RNA was reverse transcribed (as above) and fluorescently labelled with Alexa Fluor 555 for C57BL/6J and Alexa Fluor 647 for BXD strain samples (as above).

C57 versus SJL 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 C57BL/6J vs. DBA/2J experiment, but sodium tetraborate instead of sodium bicarbonate was used in the labelling buffer. Again, C57BL/6J cDNA was labelled with Alexa Fluor 555 and SJL/J with Alexa Fluor 647 for DBA/2J (Invitrogen, Mt. Waverley, Vic, Australia).

Microarray experiments

Parental experiment: For each tissue, labelled cDNA was directly compared on 6 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, Mt. Waverley, Vic, Australia). 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 as above, with one array being performed for each tissue in each BXD line, giving a total of $31 \times 3 = 93$ arrays. **C57 versus SJL experiment**: Identical arrays and processing 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 version 2 (CSIRO, Australia, texttthttp://experimental.act.cmis.csiro.au/Spot/index.php). All further data processing and statistical analyses were performed using R version 2.0.0 (Ihaka and Gentleman, 1996). Gene expression data were morph background corrected and log_2 transformed. Data for controls and the 232 replicated spots of the housekeeping gene Gapd (NM_008084) were removed prior to normalization to avoid bias. **Parental 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), and 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 sub–selected the data from genes considered to be expressed in all 3 tissues in the parental experiment and then applied quantile normalization. The log_2 ratios of intensities, $M = log_2R - log_2G$, (referred to as $M-$values) were subsequently used as expression measurements. **C57BL/2J vs. SJL/J experiment**: Processing as for parental experiment.

Differential expression in parental strains across multiple tissues

We classified genes as reliably detected if their log mean intensity, $A = 0.5(log_2 R + log_2 G)$, was greater than the 95–th percentile of negative controls present on our arrays, in all three tissues. B statistics were then calculated for all genes, using default parameters in the R limma library version 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)> 3 and an $A-$value greater than the intensity threshold. 6,075 genes were detected above in all three tissues; and of these 755 were genetically influenced in one or more tissue.

Cross-tissue correlation analysis

In order to identify the genes that have similar expression patterns to gene g_i in all tissues, we adopted a correlation-based approach. There are 3 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, that is, 755 genes×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:

$$E_{BKL} = (E_{brain}|E_{kidney}|E_{liver})$$

We then computed a $G \times G$ correlation matrix, C_{BKL}, from E_{BKL} using Spearman's ρ as a distance measure. 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}, representing an undirected simple graph. In the present study, all networks were generated using a threshold of $|\rho| \geq 0.775$ (see next section for discussion). The cross-tissue co–expression network, defined from this adjacency matrix, was visualised using custom R code using the `igraph` and `RGL` libraries. Nodes in **Figure 2** were laid out using the 2D–Fruchterman–Reingold algorithm (Fruchterman and Rheingold, 1991), computed using implementations available in the `igraph` library in R, and visualised using the `rglplot` function (see figure legends for specific details).

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 the Pearsons 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, by permutation analysis. We repeat this analysis for random CGGs, chosen by randomly sampling the same number of genes from the set of 755 genes, obtaining 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 1000 random CGGs, and count the number of times the $P-$value < 0.05, divided by the number of permutations. Similar results are obtained if the random genes are resampled from the set of 6075 genes, or if the random genes are compared to the observed CGGs centroid, rather than the random CGGs centroid (data not presented).

Inter–BXD–strain coherency

The coherency test-statistic is designed to measure how consistent the directionality of relative expression is in a set of genes (see **Results: The collective behaviour of 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, $G = \{g_1, \ldots, g_N\}$, 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:

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

where *sign* is the sign function, defined as:

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

Thus, this vertex-based-coherency score ranges from $[-1, 1]$, with values closer to $+1$ indicating more coherent up-regulated expression, values closer to -1 indicating more coherent down-regulated expression and values closer to 0 indicating less coherent expression. **Permutation test**: we chose 1000 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 1000 randomised samples, only 6 scores were observed to be larger than 0.77, then the $P-$value would be 6/1000=0.006

To test for the significance of a coherency score in just a single microarray test on a single BXD strain, we resample the appropriate number of genes in a CGG from the 755 gene set, conditional on the observed coherence of the actual CGG genes in DBA/2J–vs–C57 experiment; for example if coherence of 10 genes is +0.8 we randomly identify 8 up regulated and 2 down regulated in the DBA/2J–vs–C57 comparison and calculate the coherence of these genes in each of the 31 BXD lines, repeating the process 1000 times. We score and individual BXDs coherency score as being significant only if observed coherencies equal or greater than all 1000 random tests (or less than for $-$ve coherency).

Gene Ontology 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 > 5 genes annotated to them, either directly or via transitive relationships in the ontology. We employed 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).

Transcription factor binding motifs

The GenBank sequences for each of the 6,075 expressed genes were aligned to the ncbi35.1 build of the mouse genome using BLAT (version 32x1; Kuhn *et al.*, 2007), and the best hits were retained. The upstream 1000 bp from these sequences was then retrieved using BioPerl into FastA formatted files. Repeat regions were masked to lowercase letters using RepeatMasker (version open–3.1.6) and RepBase (version 20061006) using the following flags: "`--species mouse --xsmall --gff`". Then the upstream sequences for all of the genes in each connected component 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 (version Mar 29 2006) was run to search for over-represented motifs in the upstream sequences from the genes in each regulon compared to a background set of sequences from the 6075 expressed genes. This data was permuted 1000 times to generate $P-$values for over/under representation in the data sets. The following flags were used when running clover: "`-l -t 0.05`".

eQTL analysis for genes in CGG2

For all expression phenotypes in CGG2 (63 genes), we calculated linkage test statistics for the closest 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 CGG2. We identified the SOX binding motif as being over-represented, and since most SOX proteins are expected to recognised the same motif (P Koopman, personal communication), we consider all Sox genes. This analysis was performed in each of the three tissues separately. We estimated 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 × gene × tissue) using the Bonferroni correction.

Acknowledgements

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 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).

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.

Beck, J. A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. F., and Fisher, E. M. (2000). Genealogies of mouse inbred strains. *Nat Genet*, **24**(1), 23–25.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, **102**(5), 1572–1577.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M. T., Wiltshire, T., Su, A. I., Vellenga, E., Wang, J., Manly, K. F., Lu, L., Chesler, E. J., Alberts, R., Jansen, R. C., Williams, R. W., Cooke, M. P., and de Haan, G. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet*, **37**(3), 225–232.

Chesler, E. J., Lu, L., Wang, J., Williams, R. W., and Manly, K. F. (2004). Webqtl: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nat Neurosci*, **7**(5), 485–486.

Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., Threadgill, D. W., Manly, K. F., and Williams, R. W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*, **37**(3), 233–242.

Cotsapas, C. J., Williams, R. B. H., Pulvers, J. N., Nott, D. J., Chan, E. K. F., Cowley, M. J., and Little, P. F. R. (2006). Genetic dissection of gene regulation in multiple mouse tissues. *Mamm Genome*, **17**(6), 490–495.

Foss, E. J., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., and Kruglyak, L. (2007). Genetic basis of proteome variation in yeast. *Nat Genet*, **39**(11), 1369–1375.

Freeman, T. C., Goldovsky, L., Brosch, M., van Dongen, S., Maziere, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol*, **3**(10), 2032–2042.

Frith, M. C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004). Detection of functional dna motifs via statistical over-representation. *Nucleic Acids Res*, **32**(4), 1372–1381.

Fruchterman, T. M. and Rheingold, E. M. (1991). Force-directed placement. *Software Experience and Practice*, **21**(11), 1129–1164.

Gatti, D., Maki, A., Chesler, E. J., Kirova, R., Kosyk, O., Lu, L., Manly, K. F., Williams, R. W., Perkins, A., Langston, M. A., Threadgill, D. W., and Rusyn, I. (2007). Genome-level analysis of genetic regulation of liver gene expression networks. *Hepatology*, **46**(2), 548–557.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (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., Castellanos, R., Brozell, A., Schadt, E. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*, **2**(8), e130.

Gibson, G. and Weir, B. (2005). The quantitative genetics of transcription. *Trends Genet*, **21**(11), 616–623.

Goring, H. H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., Jowett, J. B. M., Abraham, L. J., Rainwater, D. L., Comuzzie, A. G., Mahaney, M. C., Almasy, L., MacCluer, J. W., Kissebah, A. H., Collier, G. R., Moses, E. K., and Blangero, J. (2007). Discovery of expression qtls using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*, **39**(10), 1208–1216.

Hovatta, I., Zapala, M. A., Broide, R. S., Schadt, E. E., Libiger, O., Schork, N. J., Lockhart, D. J., and Barlow, C. (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, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Muller, A., Cook, S. A., Kurtz, T. W., Whittaker, J., Pravenec, M., and Aitman, T. J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet*, **37**(3), 243–253.

Ihaka, R. and Gentleman, R. C. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Komili, S. and Silver, P. A. (2008). Coupling and coordination in gene expression processes: a systems biology view. *Nat Rev Genet*, **9**(1), 38–48.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pedersen, J. S., Hsu, F., Hinrichs, A. S., Harte, R. A., Diekhans, M., Clawson, H., Bejerano, G., Barber, G. P., Baertsch, R., Haussler, D., and Kent, W. J. (2007). The ucsc genome browser database: update 2007. *Nucleic Acids Res*, **35**(Database issue), D668–73.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., Williams, R. W., Kendziorski, C., and Attie, A. D. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet*, **2**(1), e6.

Lonnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Maciag, K., Altschuler, S. J., Slack, M. D., Krogan, N. J., Emili, A., Greenblatt, J. F., Maniatis, T., and Wu, L. F. (2006). Systems-level analyses identify extensive coupling among gene expression machines. *Mol Syst Biol*, **2**, 2006.0003.

Maniatis, T. and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*, **416**(6880), 499–506.

Nadler, J. J., Zou, F., Huang, H., Moy, S. S., Lauder, J., Crawley, J. N., Threadgill, D. W., Wright, F. A., and Magnuson, T. R. (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, W.-P., Hunt, P., Deighton, N., and Gibson, G. (2007). Quantitative trait transcripts for nicotine resistance in drosophila melanogaster. *Nat Genet*, **39**(2), 264–268.

Pavlidis, P. and Noble, W. S. (2001). Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biol*, **2**(10), RESEARCH0042.

Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., Hubner, N., and Aitman, T. J. (2006). Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet*, **2**(10), e172.

Rockman, M. V. and Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet*, **7**(11), 862–872.

Sandberg, R., Yasuda, R., Pankratz, D. G., Carter, T. A., Del Rio, J. A., Wodicka, L., Mayford, M., Lockhart, D. J., and Barlow, C. (2000). Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc Natl Acad Sci U S A*, **97**(20), 11038–11043.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Mol Syst Biol*, **3**, 88.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**, Article3.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S. E., Tavare, S., Deloukas, P., and Dermitzakis, E. T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet*, **1**(6), e78.

Taylor, B. A., Wnek, C., Kotlus, B. S., Roemer, N., MacTaggart, T., and Phillips, S. J. (1999). Genotyping new bxd recombinant inbred mouse strains and comparison of bxd and consensus maps. *Mammalian Genome*, **10**, 335–348.

Tsankov, A. M., Brown, C. R., Yu, M. C., Win, M. Z., Silver, P. A., and Casolari, J. M. (2006). Communication between levels of transcriptional control improves robustness and adaptivity. *Mol Syst Biol*, **2**, 65.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of jaspar, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, **34**(Database issue), D95–7.

Voy, B. H., Scharff, J. A., Perkins, A. D., Saxton, A. M., Borate, B., Chesler, E. J., Branstetter, L. K., and Langston, M. A. (2006). Extracting gene networks for low-dose radiation using graph theoretical algorithms. *PLoS Comput Biol*, **2**(7), e89.

West, M. A. L., Kim, K., Kliebenstein, D. J., van Leeuwen, H., Michelmore, R. W., Doerge, R. W., and St Clair, D. A. (2007). Global eqtl mapping reveals the complex genetic architecture of transcript-level variation in arabidopsis. *Genetics*, **175**(3), 1441–1450.

Williams, R. B. H., Chan, E. K. F., Cowley, M. J., and Little, P. F. R. (2007). The influence of genetic variation on gene expression. *Genome Res*, **17**(12), 1707–1716.

Yang, J. Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cdna microarray data. In M. Bittner, Y. Chen, A. Dorsel, and E. Dougherty, editors, *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE*, pages 1–12.

Yang, X., Schadt, E. E., Wang, S., Wang, H., Arnold, A. P., Ingram-Drake, L., Drake, T. A., and Lusis, A. J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res*, **16**(8), 995–1004.

Figure Legends

Figure 1: **Network properties of the 755 genes across a range of correlation thresholds.** Networks were constructed for a range of correlation thresholds from 0.05 to 1.0, and each resulting network was tested for: *(A)* the number of gene–gene correlations (edges) in the network; *(B)* the number of connected components in the network; *(C)* the median connected component size (log scale on $y-$axis); *(D)* the size (number of genes) of the largest connected component (log scale on $y-$axis); *(E)* the average degree of all vertices (log scale on $y-$axis); and *(F)* the clustering co–efficient. Within each plot, the solid black dots are the observed data points in the cross-tissue correlation network, with the 0.775 data point displayed as an open circle. 1000 network permutations were performed (see text) to generate a null distribution, which is represented as the grey area. The heavy dashed line is the mean of the null distribution.

Figure 2: *(A)*: Correlations between genes are displayed as a graph: edges connect two genes if those genes are correlated with an absolute value of Spearmans $|\rho| > 0.775$. 212 of the 755 genetically influenced genes (see text) pass this threshold and are positioned in the $x - y$ plane based on a 2–dimensional Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991). Disconnected clusters of genes (CGGs) with at least three genes in them are coloured and numbered. *(B)* panels show expression differences of genes in the relevant CGGs measured in each BXD strain in 3 tissues (1st panel brain, 2nd kidney and 3rd liver). The vertical axis is fold change vs. C57BL/6J ($M-$values) of mRNA level in each of the 31 BXD strains (horizontal axis). Each individual genes $M-$values are plotted as grey lines, with thick coloured lines representing the CGG centroids (blue, green and red for brain, kidney and liver respectively). Note the striking differences of the same genes expression patterns in the three different tissues.

Figure 3: CGGs are highly correlated to their own centroid. In each plot, the centroid for the CGG was computed, and the distributions of R^2 of each gene in the CGG to the centroid is plotted as a thick coloured line, with R^2 along the $x-$axis, and the density along the $y-$axis. The grey lines in each plot correspond to the distributions of R^2 from 1000 randomly sampled sets of genes (see **Methods**). Row 1 contains data generated from combining the gene expression data from the three single-tissues together, and rows 2–4 correspond to using the single-tissue gene expression data from brain, kidney and liver respectively. Columns 1–5 correspond to CGGs 1–5, respectively.

Figure 4: Coherency analysis. *(A)* Coherency overview: a CGG containing 12 genes is identified by correlation analysis in the 31 BXD strains; the expression ratios from a comparison of the progenitor mouse strains for each of these 12 genes are shown (most genes are up–regulated); the coherency score is calculated (see supplementary information); statistical significance is determined via permutation; 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, and for two separate pair–wise comparisons of strains (see below). *(B)* Coherency results: we plot the coherency scores for each CGG, in the brain, kidney and liver for the comparison of 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). Stars indicate the degree of statistical significance ($^* = P < 0.05$, $^{**} = P < 0.005$). *(C)*: The same data as in *(B)*, but re–ordered so that the tissues are grouped together.

Figure 5: changes in coherency across the individual BXD strains. The coherency of CGG3 in Kidney *(A)* and CGG4 in Kidney *(B)* in all 33 strains investigated in this study. The CGG depicted in *(A)* has the whole range of variable expression patterns, from completely up–regulated, to completely down–regulated. All genes in the CGG depicted in *(B)* are down–regulated in DBA/2J, but in the majority of BXD strains (all of which contain differing amounts of the DBA/2J genetic background), all genes in the same CGG are up–regulated.

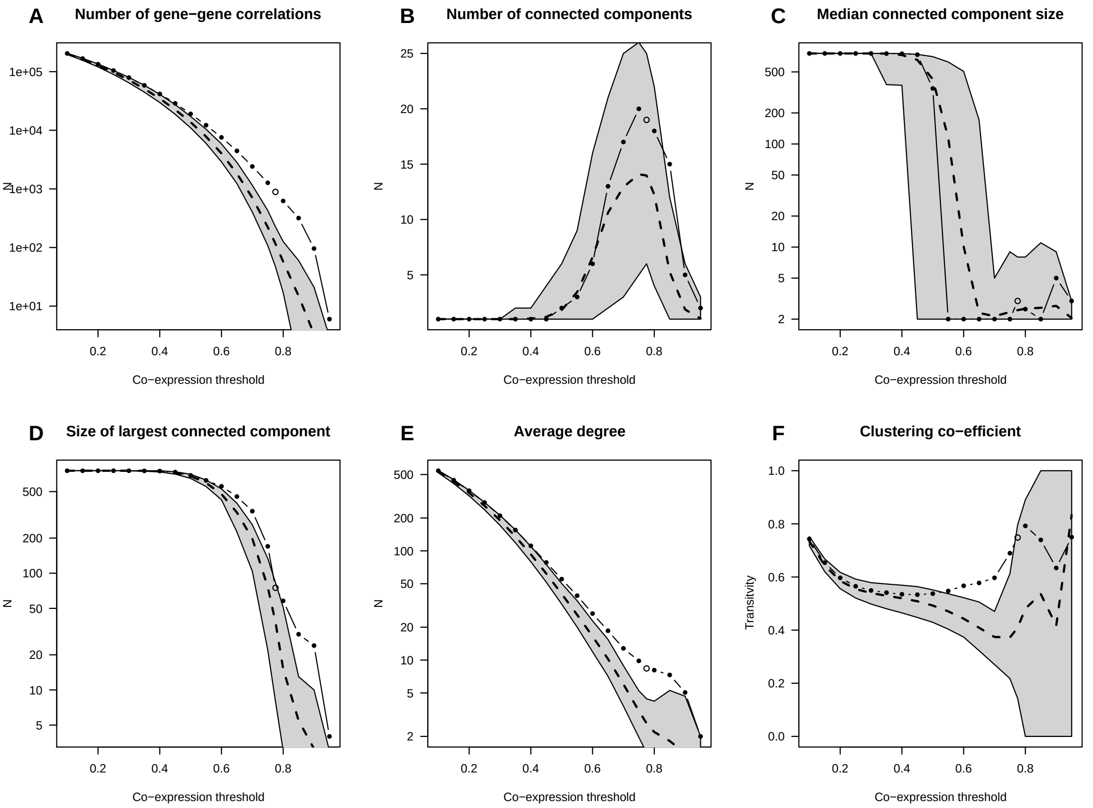

Figure 1: Cowley *et al.*

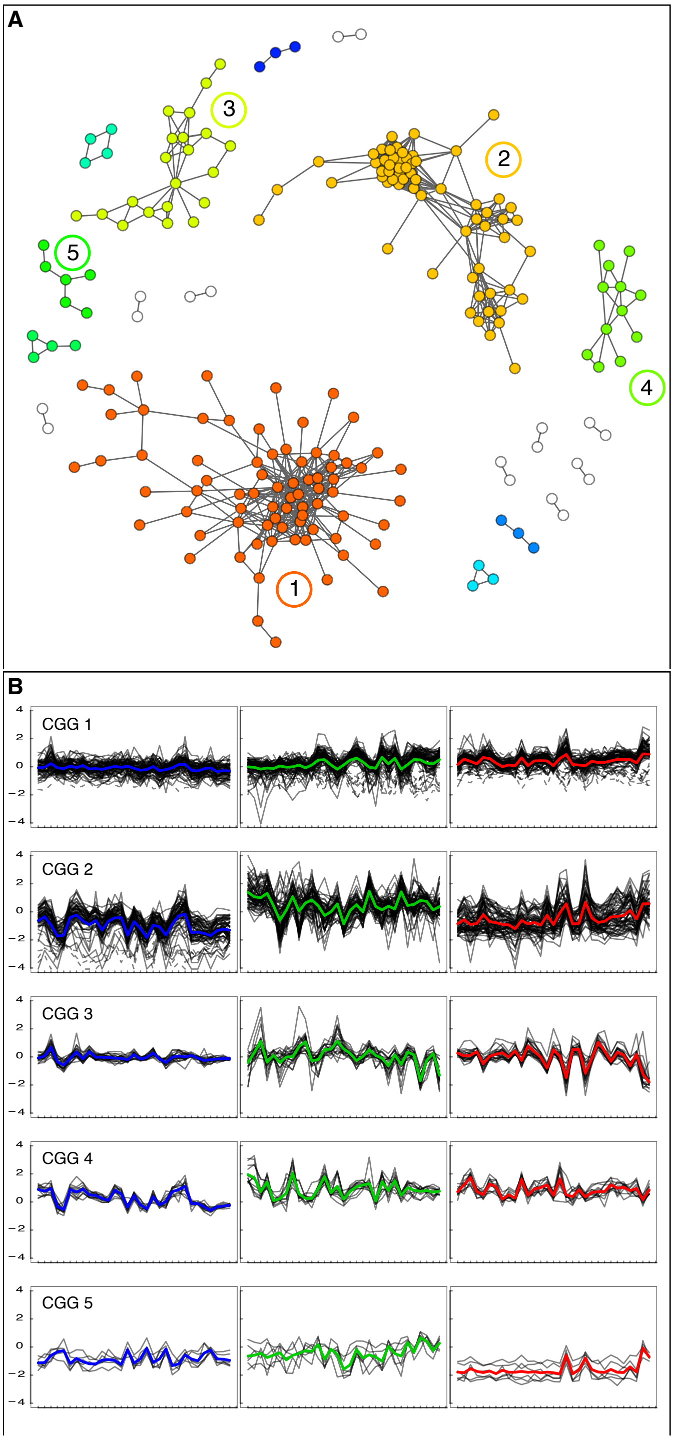

Figure 2: Cowley *et al.*

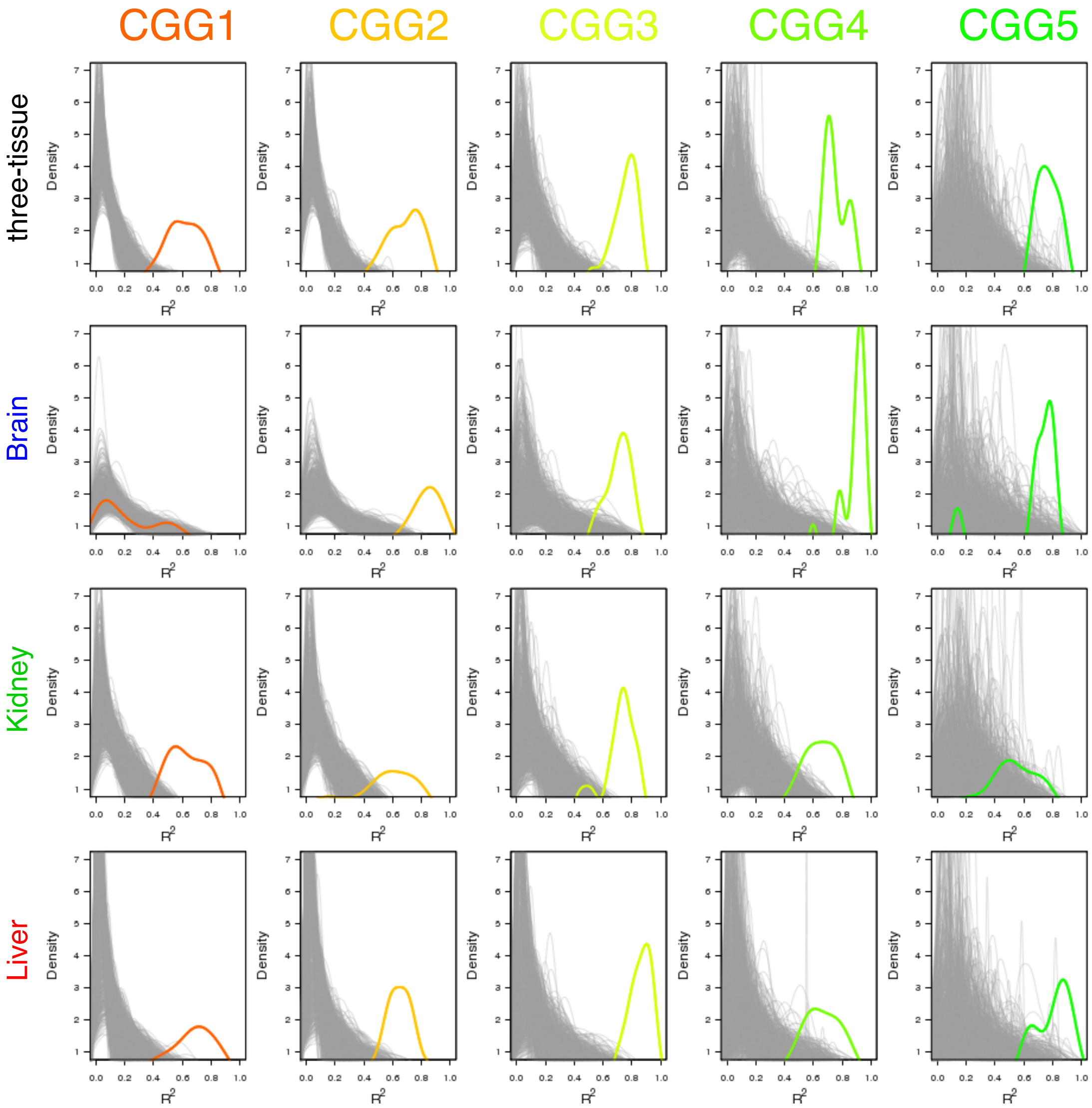

Figure 3: Cowley *et al.*

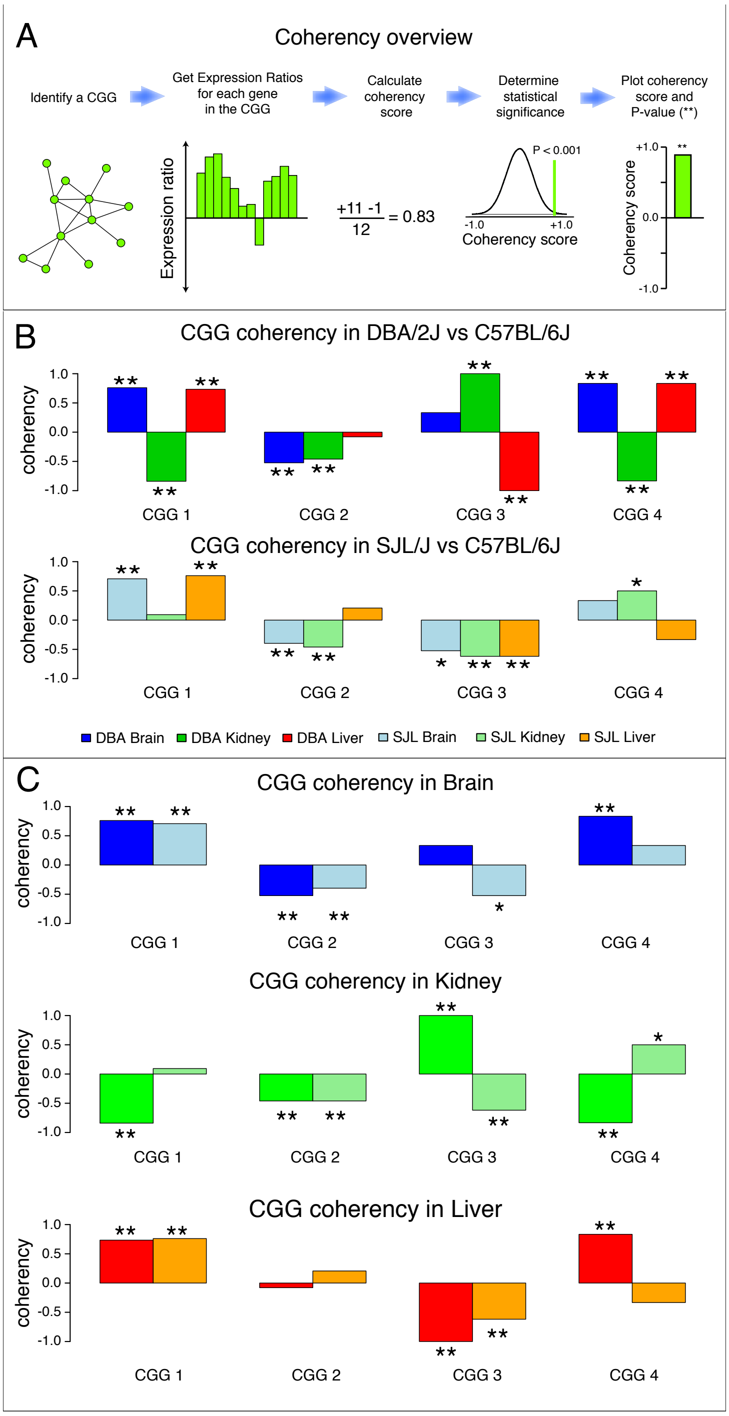

Figure 4: Cowley *et al.*

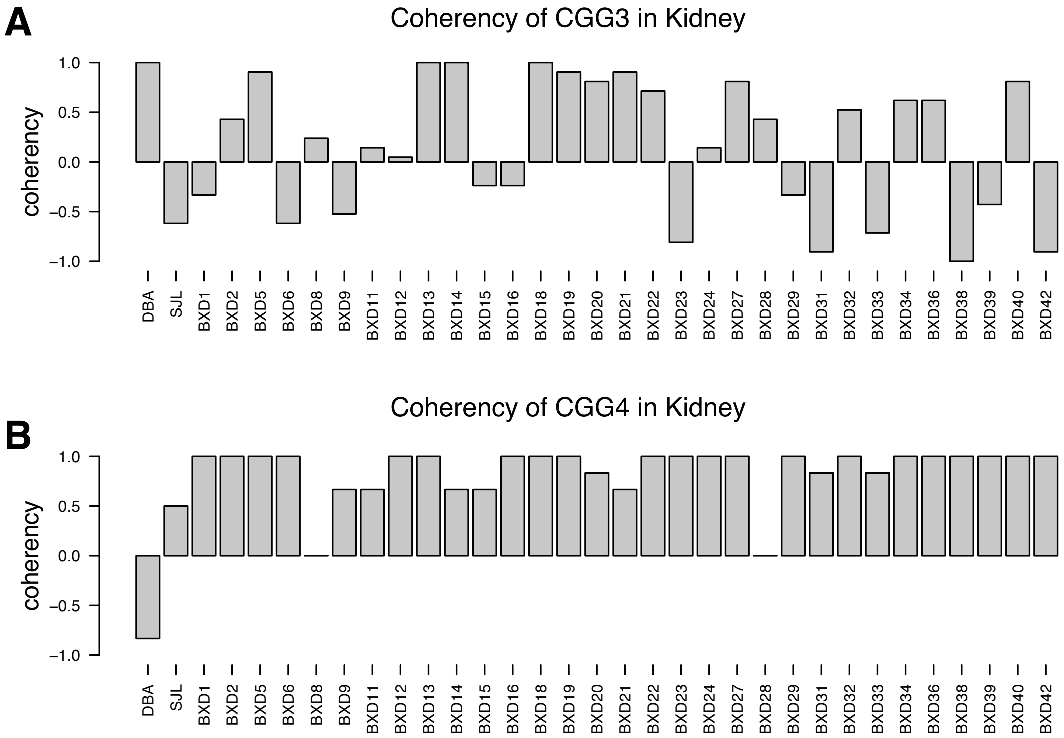

Figure 5: Cowley *et al.*