Extensive compensatory *cis-trans* regulation in the evolution of mouse gene expression

Angela Goncalves, Sarah Leigh-Brown, David Thybert, et al.

Genome Res. 2012 22: 2376-2384 originally published online August 23, 2012
Access the most recent version at doi:10.1101/gr.142281.112

Extensive compensatory *cis–trans* regulation in the evolution of mouse gene expression

Angela Goncalves,[1,5] Sarah Leigh-Brown,[2,3,5] David Thybert,[1,6] Klara Stefflova,[3,6] Ernest Turro,[2,3] Paul Flicek,[1,4] Alvis Brazma,[1] Duncan T. Odom,[2,3,4,7] and John C. Marioni[1,7]

[1]*European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom;* [2]*Department of Oncology, University of Cambridge, Hills Road, Cambridge CB2 0XZ, United Kingdom;* [3]*Cancer Research UK, Cambridge Research Institute, Robinson Way, Cambridge CB2 0RE, United Kingdom;* [4]*Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*

Gene expression levels are thought to diverge primarily via regulatory mutations in *trans* within species, and in *cis* between species. To test this hypothesis in mammals we used RNA-sequencing to measure gene expression divergence between C57BL/6J and CAST/EiJ mouse strains and allele-specific expression in their F1 progeny. We identified 535 genes with parent-of-origin specific expression patterns, although few of these showed full allelic silencing. This suggests that the number of imprinted genes in a typical mouse somatic tissue is relatively small. In the set of nonimprinted genes, 32% showed evidence of divergent expression between the two strains. Of these, 2% could be attributed purely to variants acting in *trans*, while 43% were attributable only to variants acting in *cis*. The genes with expression divergence driven by changes in *trans* showed significantly higher sequence constraint than genes where the divergence was explained by variants acting in *cis*. The remaining genes with divergent patterns of expression (55%) were regulated by a combination of variants acting in *cis* and variants acting in *trans*. Intriguingly, the changes in expression induced by the *cis* and *trans* variants were in opposite directions more frequently than expected by chance, implying that compensatory regulation to stabilize gene expression levels is widespread. We propose that expression levels of genes regulated by this mechanism are fine-tuned by *cis* variants that arise following regulatory changes in *trans*, suggesting that many *cis* variants are not the primary targets of natural selection.

[Supplemental material is available for this article.]

Identifying and characterizing the regulatory mechanisms responsible for changes in gene expression levels is a key goal of molecular biology (Stern and Orgogozo 2008). Transcriptional variation can explain phenotypic differences both between and within species—for example, differential expression of the *Tan* gene between North American *Drosophila* species underlies divergence of pigmentation (Wittkopp et al. 2009), whereas variation in the expression levels of the *LCT* gene within the human population is associated with lactose tolerance (Tishkoff et al. 2007; Majewski and Pastinen 2011).

All regulatory mutations that alter the expression level of a gene can be classified according to their location and their linkage disequilibrium relative to the gene that they affect. For a given gene, its expression level can diverge between or within populations due to (1) regulatory mutations acting in *trans* to that gene, which mediate differential expression via a diffusible element such as a protein or ribonucleic acid (e.g., a change in the expression level of an upstream transcription factor), or (2) regulatory mutations acting in *cis* to that gene, which mediate differential expression directly by altering the local genomic sequence (e.g., a mutation in the promoter sequence that alters a transcription factor binding site). This distinction reflects underlying differences in the inheritance of the change in gene expression levels and the resulting selective pressures to which a mutation is exposed (Wray 2007; Lemos et al. 2008; McManus et al. 2010).

Given the different evolutionary implications of these regulatory mechanisms, a significant amount of effort has been expended on investigating the contribution of regulatory changes in *cis* and in *trans* to divergence in gene expression. This has been studied at the single gene level using enhancer swap experiments, where orthologous regulatory sequences from two species are used to drive the expression of a reporter gene in one of the species under study (Gordon and Ruvinsky 2012). At the genome-wide level, regulatory mutations can be identified using expression quantitative trait loci (eQTL) studies (Majewski and Pastinen 2011). In a typical eQTL study, the expression levels of all genes are measured across a population, and genetic variants (e.g., single nucleotide polymorphisms, or SNPs) are typed in the same set of samples. For every gene, expression levels are correlated with the genotypes measured at each SNP and a significant SNP–gene association suggests that a regulatory mutation affecting the gene's expression is in high linkage disequilibrium with the SNP identified.

Despite their popularity, a significant challenge faced by eQTL studies is to distinguish between regulatory divergence in *cis* and in *trans* (Pickrell et al. 2010). In particular, eQTL studies that test all SNPs against all genes are statistically underpowered for identifying variants. To overcome this problem, eQTL studies typically focus on identifying "*cis*-eQTL" by concentrating only on

[5]These authors contributed equally to this work.
[6]These authors contributed equally to this work.
[7]Corresponding authors
E-mail marioni@ebi.ac.uk
E-mail Duncan.Odom@cancer.org.uk
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.142281.112. Freely available online through the *Genome Research* Open Access option.

22:2376–2384 © 2012, Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

SNPs that are located proximal to a gene (for example within 200 kb of its transcription start site). This restriction removes the possibility of identifying regulatory variants distal to a gene, which are most likely to regulate the gene's expression in *trans* (Gibson and Weir 2005). Further, while there is evidence that many *cis*-eQTL do regulate the proximal gene's expression in *cis* (Pickrell et al. 2010), this is not always the case (Doss et al. 2005).

A more powerful method for studying transcriptional divergence is to compare gene expression differences between first generation (F1) hybrids of two homozygous parents (F0s), such as inbred laboratory lines of mice, fruitfly, or yeast (Wittkopp et al. 2004; Tirosh et al. 2009). In F1 hybrids, coding variants between the two parents allow allele-specific expression to be measured. In the first generation hybrid of inbred strains, regulatory variants acting in *cis* remain linked to their target gene and result in allele-specific expression. Regulatory mutations acting in *trans*, however, influence both parental alleles equally in the F1 hybrid, since they are in the same nuclear environment. A comparison of differential expression between the parent strains (F0) and allele-specific expression in the hybrid (F1) therefore distinguishes regulatory divergence in *cis* and regulatory divergence in *trans* across the entire transcriptome. For genes with differential expression between the two parental strains, if the ratio of allele-specific expression is equal to the ratio of expression between the parent strains, the difference can be attributed to one or more regulatory variants acting in *cis*. In contrast, if both alleles are expressed at the same level in the F1 hybrids, the difference is due to one or more regulatory variants that act in *trans*.

RNA-sequencing (RNA-seq) has been used to measure expression levels in the F0 animals/strains and allele-specific expression (ASE) in the F1 hybrids of *Drosophila* and of yeast populations (Emerson et al. 2010; McManus et al. 2010). It has also been used to study parent-of-origin effects between mouse strains (Gregg et al. 2010; Xie et al. 2012). However, the regulation of gene expression in a mammalian system using an F1 hybrid model has remained mostly unexplored. The increased genome size and complexity in mammals relative to *Drosophila* and yeast results in a considerably greater number of sites where mutations can arise either in *cis* or in *trans*. Here, we used RNA-seq to measure transcript abundance in liver samples taken from multiple mice from two inbred mouse strains and their F1 hybrids. Providing insight into an issue that has recently proved contentious, we identified hundreds of genes with parent-of-origin specific patterns of expression. More importantly, these data allowed us to investigate whether regulatory divergence in *trans* plays a major role in explaining differences in gene expression levels, as has recently been observed in *Drosophila* (McManus et al. 2010). Contrary to this, we found that a combination of *cis* and *trans* acting variants drives the divergence of gene expression levels in closely related mammals.

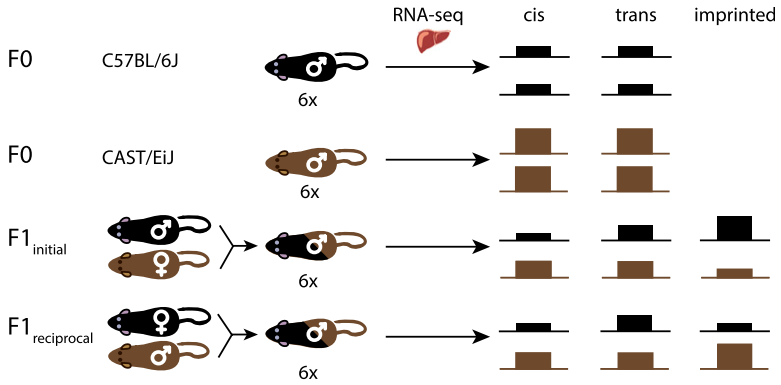

Figure 1. Study design. Liver samples were collected from six adult male mice from each of four groups: C57BL/6J, CAST/EiJ, F1 initial cross hybrid of a C57BL/6J male with a CAST/EiJ female, and F1 reciprocal cross hybrid of a C57BL/6J female with a CAST/EiJ male. For each sample, the polyadenylated fraction of total RNA was sequenced on an Illumina GAIIx with 72-bp paired-end reads.

Results

Two inbred mouse strains, C57BL/6J and CAST/EiJ, were crossed to generate both initial and reciprocal F1 crosses (Fig. 1). These strains were derived from different subspecies of *Mus musculus* with a divergence time of approximately 1 million years. Six biological replicates of liver gene expression were generated for each genetically distinct class of mice (F0 C57BL/6J, F0 CAST/EiJ, F1i–C57BL/6J × CAST/EiJ, F1r–CAST/EiJ × C57BL/6J, where the male parent is listed first). For each class, samples were collected from a single lobe of the liver from six male mice between the ages of 4 and 6 mo. These 24 samples were then processed to generate strand-specific RNA-seq libraries, which were sequenced on the Illumina GAII platform using 72-bp paired-end reads. Sequenced reads were mapped to the appropriate transcriptome using Bowtie (Langmead et al. 2009) and gene expression estimates were obtained using MMSEQ (Turro et al. 2011; see Methods). Our analysis pipeline was assessed using a number of metrics, including the ability to measure expression levels accurately in an artificial F1 library created by combining reads from two F0 libraries, one from each parental strain (Methods; Supplemental Figs. 1, 2). When comparing the gene expression in the F0s with the allelic expression in the in silico F1 library (Supplemental Fig. 2), we found a very good agreement between the two (Pearson correlation ≥0.97). A high correlation was also observed when we examined the correlation between the fold change of the F1 alleles and the fold change in the F0 parents. Both of these observations give us confidence in our bioinformatics pipeline.

The power of our approach is determined by the number of genes expressed in liver that contain a genetic variant between the parents, thus allowing the two alleles to be distinguished in the F1 hybrids. Of the set of 36,229 mouse genes defined in the Ensembl database (version 59; Flicek et al. 2012), we detected the expression of 13,551 (37%) genes in the F0 mice, and 11,183 (31%) genes in both the F0 and the F1 mice (Methods). Of this latter set, 10,909 (98%) contain at least one single nucleotide polymorphism (a SNP) between the two parental strains. We validated our RNA-seq derived measures of allele-specific expression by applying pyrosequencing to a set of five genes, three of which contain multiple SNPs (Supplemental Fig. 3; Methods).

Approximately a quarter of genes are differentially expressed between C57BL/6J and CAST/EiJ and circadian rhythm contributes to gene expression variation between biological replicates

To characterize the divergence of gene expression levels between C57BL/6J and CAST/EiJ we considered the set of 13,551 genes

expressed in the F0 mice, and used DESeq (Anders and Huber 2010) to identify genes that are differentially expressed. At a false discovery rate (FDR) cutoff of 5%, 3906 genes (29%) were identified as differentially expressed, with 1940 (49.6%) being more highly expressed in C57BL/6J (Supplemental Table 1). Using GeneTrail (Backes et al. 2007) we determined that the genes up-regulated in C57BL/6J were significantly enriched for genes involved in fatty acid metabolism (GO category "Peroxisome") (Supplemental Table 2). Conversely, genes involved in drug metabolism (KEGG category "Drug metabolism") were highly enriched in the set that was up-regulated in CAST/EiJ compared with C57BL/6J (Supplemental Table 3).

Since we had six biological replicates in each genetic class, we were able to identify genes with expression levels that were highly variable among individuals with the same genetic background. Within each strain we identified the top 10% of variable genes using a dispersion metric adapted from Anders and Huber (2010) (Methods; Supplemental Fig. 4). The 423 genes that were highly variable in both C57BL/6J and CAST/EiJ were significantly enriched in genes that play a role in the regulation of circadian rhythm (e.g., *Dbp*, *Npas2*, *Arntl*, *Per1*, *Per3*). This set also includes a number of genes with expression levels that have been shown to vary in response to external stimuli such as fed or fasted state (*Egr1*) and injury/infection (*Cish*). We identified a small number of highly variable genes, *Hba-a1*, *Hba-a2*, *Hbb-b2*, and *Saa2*, which are not endogenously expressed in liver cells (K Stefflova and DT Odom, unpubl.) and are instead likely expressed at high levels in a minority of samples due to blood contamination during sample processing (Supplemental Table 4).

The identification of imprinted genes is strengthened by multiple replicates

Imprinting in mammals describes the situation where the allele-specific expression of a gene is determined purely by the sex of the parent from whom the allele is inherited. Mechanistically, most imprinting is thought to arise via differential methylation during gametogenesis (Li and Sasaki 2011). Imprinting can be a confounding factor in an F1 hybrid study design, as it results in allele-specific expression independent of regulatory divergence between the parental strains; by identifying imprinted genes we can remove them from downstream analyses. The extent of imprinting in the mouse has also become contentious, with reports indicating low hundreds (DeVeale et al. 2012) to thousands (Gregg et al. 2010) of imprinted loci in the developing mouse brain of the same genetic cross we report here. Our data allowed us to independently estimate the extent of imprinting in mouse somatic tissues.

To determine the extent of imprinting in mammalian liver cells, we compared allele-specific expression in the initial (F1i) and the reciprocal (F1r) hybrid crosses. We used a model based upon the Beta-Binomial distribution to find genes where the maternal allele is significantly more expressed than the paternal allele in both the initial and reciprocal crosses—these genes are likely to be enriched for those that are paternally imprinted (Methods). Analogously, we found genes that are maternally imprinted. After excluding genes on the X chromosome, 290 and 245 genes showed evidence of being paternally and maternally imprinted, respectively, corresponding to 5% of all expressed genes containing a genetic variant (Fig. 2; Supplemental Table 5). Many of these genes are in distinct genomic clusters that have previously been associated with genetic imprinting, including the Callipyge locus on chromosome 12 (*Meg3*) and the *Igf2r* cluster (*Slc22a3*, *Igf2r*, *Mas1*). We further compared the set of imprinted genes identified in our study to a set of 20 genes imprinted in the mouse liver as described in two earlier studies (Haig 2004; Lees-Murdock and Walsh 2008). Of the 20 genes identified previously, we were able to assess the imprinting status of eight that were expressed in adult liver samples and that contained at least one SNP between C57BL/6J and CAST/EiJ. We identified seven of these genes as being imprinted (the only gene that was not identified, *Airn*, had low and variable expression across the samples in our data set).

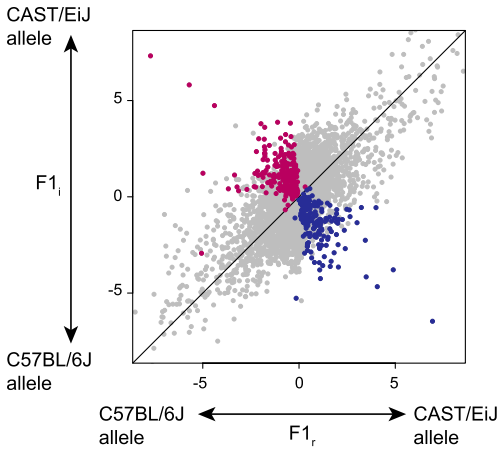

Figure 2. Imprinted genes. After removing genes on the sex chromosomes, similar numbers of genes were found to be expressed from the maternal allele (290 genes, colored pink) and from the paternal allele (245 genes, colored blue). The average $\log_2$ expression fold change between the two alleles in the initial cross hybrids (F1i) and between the two alleles in the reciprocal cross hybrids (F1r) is plotted on the *y*- and *x*-axis, respectively.

Gregg and colleagues recently tested for parent-of-origin effects genome-wide in brain using a C57BL/6J–CAST/EiJ F1 hybrid system, and identified 1308 candidate imprinted genes (Gregg et al. 2010). Of this set, 534 are also expressed in the liver, of which we classify 25 (5%) as imprinted in our data. Possible reasons for the small overlap are the tissue specificity of some imprinted loci and the important role of imprinting in brain development. Two differences in analytical approach can also in part explain this disparity. First, our study uses six replicates for the initial and reciprocal cross hybrids, lending it greater specificity and sensitivity than the Gregg and colleagues study in which replication was not used. Second, we assessed allelic imbalances at the gene level, while in the Gregg and colleagues study the allelic imbalances were assessed for individual SNPs. This can lead to difficulties when combining results across SNPs from the same gene, particularly when different SNPs yield contradictory results (see Turro et al. 2011 for a comparison of both approaches). Supporting this hypothesis, DeVeale et al. (2012) recently reanalyzed the data generated by Gregg and colleagues and found that the majority of the novel imprinted loci reported were either false positives or had an effect size that was too small to validate with pyrosequencing.

Most gene expression divergence is caused by a combination of regulatory variants in *cis* and in *trans*

We examined the divergence of steady-state gene expression levels using the set of 10,090 nonimprinted, autosomal genes expressed both in the F0 and F1 mice. For each gene, we used a statistical

framework based upon the Negative Binomial and Beta-Binomial distributions to assess whether the expression values in the F0 and the expression ratios in the F1 were consistent with regulatory divergence (Fig. 3A). Specifically, we looked for: (1) genes whose regulation is conserved between the two strains—these genes show no evidence of differential expression in the F0 mice and equal expression of the C57BL/6J and CAST/EiJ alleles in the F1 mice; (2) genes that show evidence of expression divergence due to one or more regulatory variant in *cis*—these genes show evidence of differential expression in the F0 mice and a concordant ratio of allele-specific expression in the F1 mice; (3) genes with expression levels that are consistent with divergence due to one or more regulatory variant in *trans*—these genes are differentially expressed in the F0 mice but show equal expression of each allele in the F1 hybrids; (4) genes that are expressed in a manner consistent with expression divergence due to one or more regulatory variant in *cis* together with one or more regulatory variant in *trans* (Methods).

In total, across the set of 10,090 genes, the majority (6872; 68%) showed evidence of their expression being regulated in a conserved fashion (Fig. 3A; Supplemental Table 6), consistent with the number of genes not differentially expressed between the parental strains (71%). The second largest class corresponds to genes with expression levels consistent with regulatory variants in *cis* acting alongside regulatory variants in *trans* (1758; 17%), with the third class corresponding to regulatory variants only in *cis* (1391; 14%). In contrast, the number of genes whose expression levels are consistent with divergence due to regulatory variation solely in *trans* is small—only 69 genes (<1%) fall into this category. When the fold change between the parental alleles was small we had slightly less power to allocate genes to the *trans* class relative to the *cis* or *cis*-and-*trans* classes (Supplemental Fig. 5). However, the difference in power is small and does not affect any of our conclusions. Additionally, the estimates of allelic expression for genes in the *trans* and conserved classes were slightly more variable than the corresponding estimates for genes in the other classes (Supplemental Table 7; Supplemental Fig. 6). This might lead to an increase in the number of false positives in these two classes; however, this does not challenge our observation that only a small number of genes are regulated purely in *trans*. When we examined the set of genes classified as being regulated in *cis* we observed a depletion of genes involved in core cellular processes (e.g., transcription or splicing; Supplemental Table 8), while the small set of genes regulated in *trans* were marginally significantly enriched for genes involved in transcription factor activity (Supplemental Table 9).

We found that a surprisingly large proportion of genes have expression levels shaped by multiple regulatory variants both in *cis* and in *trans* (Fig. 3A). To explore this category further, we subdivided it into four classes, similarly to previous studies (Landry et al. 2005). This enables regulatory variants that act in *cis* and in *trans* to be classified into sets such that the variants: (1) act in the same direction with a stronger effect from the ones in *trans* (*cis* + *TRANS*); (2) act in the same direction with a stronger effect from the ones in *cis* (*CIS* + *trans*); (3) act in opposing directions with the effect from variant(s) in *cis* being stronger (*CIS* − *trans*); (4) act in opposing directions with the effect from variant(s) in *trans* being stronger (*cis* − *TRANS*) (Methods; Fig 3B). From Figure 3B we observe an excess of opposite direction effects (categories 3 and 4; Supplemental Fig. 7, *P*-value < 10^{-16} Fisher's Exact Test), where the regulatory variants in *cis* and in *trans* act in opposite directions in the F1 hybrid.

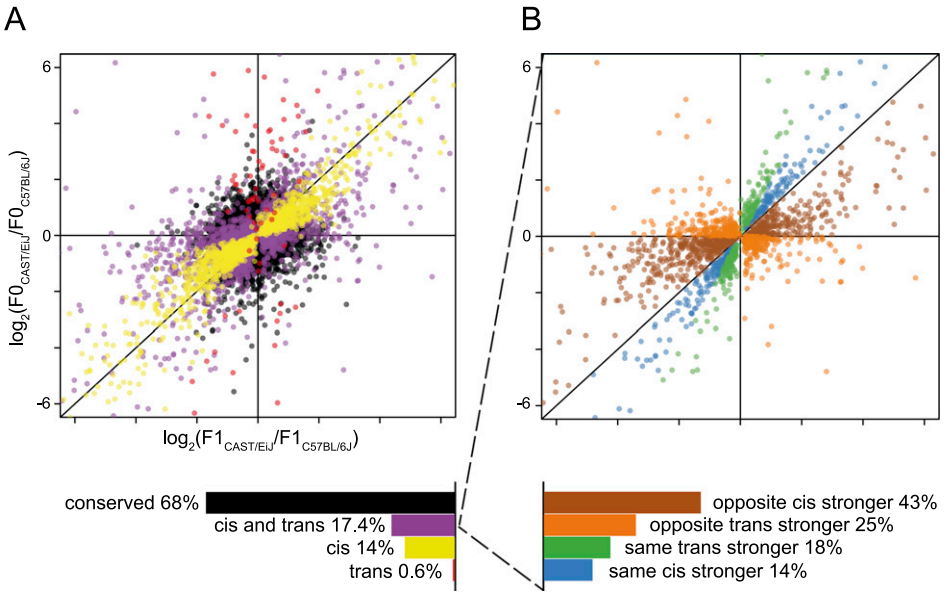

Figure 3. Classification of genes according to their pattern of gene expression divergence. The average $\log_2$ expression fold change between the alleles in the hybrids (F1) and between the parental strains (F0s) is plotted on the *x*- and *y*-axis, respectively. (*A*) Genes for which the expression levels have not diverged between the two strains are classified as conserved (colored black), while genes in which expression has diverged are classified as *cis*, *trans*, or *cis* and *trans* according to whether the divergence is explained by at least one regulatory variant acting in *cis* (colored yellow) or in *trans* (colored red), or by at least two regulatory variants, one in *cis* and one in *trans* (colored purple). (*B*) Subdivision of the *cis* and *trans* category. The regulatory variants can cause gene expression changes in the same direction with the regulatory variant in *cis* having a stronger effect on expression change than the regulatory variant in *trans* (blue), or the variant in *trans* having a stronger effect than the variant in *cis* (green). Expression changes can also be in opposite directions with the variant in *cis* having a stronger effect than the variant in *trans* (brown), or the variant in *trans* having a stronger effect than the variant in *cis* (orange).

Genes with regulatory divergence in *trans* show stronger sequence constraint

To understand if there is an association between regulatory change and sequence evolution in mammals, we examined whether there was a difference in conservation of the coding sequence for genes with divergent expression due to regulatory change purely in *cis* and regulatory change purely in *trans*. For each exon, we computed its GERP (Genomic Evolutionary Profiling) score using all mammalian species in the Ensembl compara database (Cooper et al. 2005; Vilella et al. 2009), and then looked at the distribution of conservation rates (the rate of bases with a GERP score greater than 1.4) in each regulatory category. Genes with conserved regulation and those with divergent expression driven only by diffusible element(s) in *trans* are significantly more conserved at the sequence level than genes with divergent expression that is either partially or entirely regulated by variants in *cis* (Fig. 4A; Supplemental Table 10). Importantly, although expression estimates for genes regulated by variants in *trans* were less well estimated than those with variants in *cis*, we did not find a relationship between the expression estimates' standard errors and the conservation rates (Supplemental Fig. 8). To further test this pattern, we again subdivided the set of genes that showed divergent expression due to the combined effect of regulatory variants in *cis* and in *trans* (Fig. 4B). When the two regulatory mechanisms act in concert, the genes for which the regulatory change(s) in *trans* are stronger are more conserved at the sequence level than the set of genes for which the regulatory change(s) in *cis* are stronger (Supplemental Table 10). This provides evidence that, between closely related mammalian subspecies, genes with divergent expression due to regulatory variants in *cis* have less conserved coding sequence throughout the mammalian clade than genes that have conserved regulation or that have divergent expression due to a regulatory change in a diffusible element.

Discussion

To investigate the divergence of gene regulation in mammals we tested the relative contribution of regulatory divergence only in *cis*, only in *trans*, and the action of changes both in *cis* and in *trans* over 1 million years of subspecies divergence using an F1 hybrid system. Since our approach requires allele-specific expression to be measured, our conclusions are based upon the set of genes that have at least one genetic variant between C57BL/6J and CAST/EiJ. However, since 98% of genes expressed in the mouse liver have at least one such variant, our results can be extrapolated to the entire mouse genome. We used liver tissue in this study since it is extremely homogeneous, with 70% of the cells in the liver being hepatocytes (Schrem et al. 2002). Moreover, there is no evidence that liver gene expression diverges between mammalian species at a faster rate than other tissues (Brawand et al. 2011), suggesting that the liver is representative of most somatic tissues.

Phenotypic diversity and intra-species heterogeneity in expression

Our analysis of gene expression between the two parental strains revealed a substantial number of differentially expressed genes. Despite the parent subspecies last sharing a common ancestor less than 1 million years ago, 3906 genes were differentially expressed, compared with the 3335 genes that were recently found to be differentially expressed between human and chimpanzee liver samples, two species that diverged ~6 million years ago (Blekhman et al. 2010). Our observation of greater transcriptional divergence among rodents may be in part due to our use of inbred individuals reared in the same environment and fed the same diet, thereby reducing intra-species variation in expression and increasing the statistical power to detect small, reproducible changes in expression. Nevertheless, the large number of differentially expressed

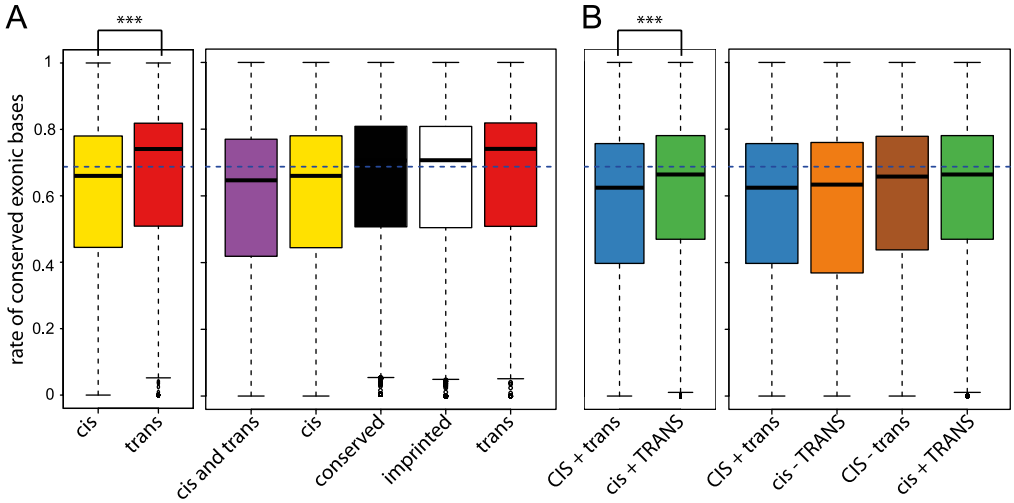

Figure 4. Exonic sequence conservation scores for the different classes of regulatory divergence. GERP conservation scores relative to all mammalian species in the Ensembl compara database were calculated for every exonic base. The proportion of bases above a GERP score of 1.4 in each exon was calculated for exons in each category. The mean conservation score for all exons is represented as a horizontal dashed blue line. (*A*) The conservation proportions for exons in the *trans* category are significantly higher than for genes in the *cis* category (*P*-value 9.7×10^{-7}; *t*-test). Imprinted and conserved genes are also significantly more conserved than the *cis* and the *cis* and *trans* categories (Supplemental Table 10). (*B*) The *cis* and *trans* category is subdivided into four subcategories: *cis* and *trans* in the same direction with *cis* stronger (*CIS + trans*), *cis* and *trans* in the same direction with *trans* stronger (*cis + TRANS*), *cis* and *trans* in opposite directions with *cis* stronger (*CIS − trans*), and *cis* and *trans* in opposite directions with *trans* stronger (*cis − TRANS*). As in *A*, for the two categories where the *cis* and *trans* regulatory variants act in concert, the set of exons from genes for which the *trans* effect is stronger also show higher conservation than the set for which the *cis* effect is stronger (*P*-value 6.8×10^{-9}; *t*-test). Supplemental Figures 10 and 11 show that the results do not change when different GERP conservation thresholds are used, or when promoter regions are considered.

genes suggests that there has been a relatively rapid divergence in the regulation of gene expression levels in the liver between C57BL/6J and CAST/EiJ compared with humans and chimpanzees. This may reflect both strong selective breeding to create the mouse strains, as well as the relatively large population size and short generation time of *Mus musculus* subspecies, compared with primates.

Many of the genes that showed high variability within both the C57BL/6J and the CAST/EiJ sample groups were linked to external stimuli, such as nutritional state (fat, glucose), chemical stimuli (pheromones, organic substances), injury/infection (cytokine signaling, inflammation), or circadian rhythm. Given the large number of biological replicates, these highly variable genes are good candidates for explaining subtle phenotypic differences between inbred mice strains and species reared within the same environment.

A continuum of imprinting

We used the F1 hybrids to investigate the extent of imprinting in the mouse liver. Our analysis suggested that a relatively small number of genes (~5% of those testable) showed statistically significant evidence of being imprinted in a representative somatic tissue. One limitation of using a reciprocal cross system to identify imprinted genes is that the two crosses have different Mitochondrial DNA and, since we are using male mice, different sex chromosomes. Hence, some of the genes we identify as imprinted might reflect these differences in the *trans* environment. Arguing against this being a significant problem, the number of imprinted genes identified and the approximately equal number of maternally and paternally imprinted genes (46% and 54%, respectively) is broadly consistent with other recent RNA-seq based studies of imprinted genes in mammalian systems (Babak et al. 2008; Wang et al. 2011).

Amongst the set of maternally imprinted genes, genes involved in cell–cell signaling, limb morphogenesis, and reproductive processes were enriched, none of which are normal functions of the liver. One possible explanation for this is that the imprinting of these genes is functionally relevant in the reproductive organs, or during morphogenesis, where parent-of-origin effects are known to play a key role, and that the imprinted status in the adult liver is a passenger effect with limited relevance to other somatic tissues. This observation is consistent with imprinting playing a minor or inconsequential role in adult mouse liver, which suggests that these genes are likely to be imprinted across all somatic tissues of the mouse.

Across imprinted genes a continuum of parent-of-origin effects was observed, from small allelic ratios to large ones. This is consistent with recent RNA-seq based studies of imprinted genes in the mouse placenta at E17.5 (Wang et al. 2011) and in the mouse brain (Gregg et al. 2010; DeVeale et al. 2012), all of which observed that only a small proportion of imprinted genes had one allele completely silenced. The increasing evidence for a continuum of parent-of-origin effects highlights a limitation of current models of how imprinting arises, and suggests further work is necessary to understand the mechanisms by which genes are partially biased toward a parental allele in the absence of complete allelic silencing.

Using the hybrid system to study the divergence of gene expression levels

Using our hybrid system we determined that the regulation of 32% of genes expressed in the mouse liver has diverged between C57BL/6J and CAST/EiJ, which is a relatively large proportion given the liver's highly conserved function and phenotype. For the small set of genes that have a regulatory variant solely in *trans*, we observed that their exonic and promoter sequence is significantly more conserved among mammals than that of genes with regulatory variants only in *cis* or in *cis* and in *trans*. One possible explanation is that these highly conserved genes are more resistant to gradual genetic divergence but remain vulnerable to rapid changes in upstream regulators. For each gene, we need at least one variant between the transcribed sequences of the parents to distinguish between the two alleles in the F1s. Given the relatively higher conservation of genes with differential expression regulated by a *trans* variant, the set of genes without a variant might disproportionately contain genes that have a regulatory variant that acts in *trans*. However, since <2% of expressed genes in the mouse liver have no coding sequence variant this is not likely to significantly affect our results.

The proportions of genes allocated to each regulatory class are consistent with a number of recent studies. When a recent eQTL study within a human population correlated all SNPs with all genes (i.e., not only focusing on *cis*-eQTL), the number of genes demonstrating divergence in expression due to a change in an element distal to the gene (likely corresponding to a *trans* mutation) was found to be extremely small (Pickrell et al. 2010). Our classification is also consistent with a recent survey of gene regulation experiments performed by manipulating the promoter region in insects and worms (Gordon and Ruvinsky 2012). By utilizing a highly curated set of genes, this study concluded that the primary form of regulatory divergence was driven by changes that arose purely in *cis* or by a combination of changes in *cis* and in *trans*, with the number of genes with only *trans* regulatory divergence being small, especially in insects. In contrast, another study of F1 hybrids using RNA-seq in *Drosophila* (McManus et al. 2010) found a larger number of genes with regulatory changes only in *trans* than only in *cis*. One explanation for this discrepancy could be genuine differences in the proportion of genes regulated in *cis* and in *trans* between mammals and flies. However, arguing against this, previous studies in *Drosophila* using a small set of genes (Wittkopp et al. 2004) found patterns of regulatory changes similar to those that we identified in the mouse. Other possible reasons include (1) differences in the length of intergenic (i.e., potentially regulatory) regions between the two taxa, (2) alternative analysis strategies, or (3) differences in the study design (McManus and colleagues used a pooled F1 hybrid approach and had only a small number of biological replicates in each set; moreover, they pooled tissue from the entire animal while we focused on samples taken from an individual tissue).

One of our most interesting observations is that the majority of genes with differences in expression levels between closely related mammals have regulatory variants both in *cis* and in *trans*. If the regulatory variants in *cis* and in *trans* arose independently in each strain, this could contribute to hybrid incompatibilities, which have been described extensively in *Drosophila* (Landry et al. 2005; McManus et al. 2010). However, if this were the explanation for most genes, we would expect to observe equal proportions of regulatory changes acting in the same and opposing directions upon gene expression levels. Instead, we observed that a significantly higher proportion of genes were regulated by *cis* and *trans* mutations that compensate for one another by acting in opposing directions. This is consistent with the action of stabilizing selection on gene expression levels and suggests that, for the majority of genes, the regulatory variants that act in *cis* and in *trans* arose on the same lineage.

The presence of at least two opposing regulatory variants could arise via an initial regulatory change in *cis* that alters the linked gene's expression, followed by a counteracting regulatory mutation in *trans*. Since changes in *trans* will likely affect a large number of genes (due to their inherently greater pleiotropy than changes in *cis*) as well as the specific gene with the *cis*-regulatory variant, this scenario is unlikely, unless all of the genes targeted by the diffusible factor have *cis*-regulatory variants that act in the same direction (relative to the change in *trans*).

The opposite order of events, where the first regulatory change arises in a diffusible element that acts in *trans* to a number of genes, seems more plausible. A regulatory variant that acts in *trans* can rapidly alter the expression profile of a large number of genes (potentially conferring a fitness advantage) but does not alone allow the fine-tuning of individual gene expression levels. Hence, the genes regulated by the specific *trans* factor may come under selective pressure to modulate their expression levels to compensate for the change imposed by the *trans* variant. The easiest way to do this is for each gene to accumulate compensatory variants that act in *cis*. This model of gene regulation evolution might be especially pertinent for domesticated animals, due to the strong selective pressure imposed to obtain specific phenotypes. Gene regulation in small wild populations might evolve under a more neutral evolutionary model. Regardless, a model where changes in gene expression levels are driven by variants that act in *trans* followed by *cis* compensation suggests that *trans* acting variants are a unique and interesting form of standing genetic variation in interbreeding mammalian populations.

In summary, our study provides a comprehensive characterization of gene regulation in closely related mouse strains and establishes the relationship between gene sequence divergence and regulatory divergence in mammals. It demonstrates that amongst the set of genes with divergent regulation between two closely related mouse strains, most are regulated by variants that have arisen both in *cis* and in *trans*. Further, most of these multiple regulatory variants act in opposing directions, suggesting extensive compensatory regulation of gene expression levels. This has important implications for understanding mammalian gene expression divergence and for understanding how speciation occurs.

Methods

Animal housing and handling

Mice were housed in the Cambridge Research Institute Biological Resource Unit with a 12-h light/dark cycle, and were provided with chow food from Lab Diet plus water ad libitum. Mice were sacrificed by cervical dislocation at 4–6 mo of age, between the hours of 9:30am and 11:30am. Each mouse was perfused with PBS and the liver removed. A sample of 50–100 mg was taken from a single liver lobe and frozen in liquid nitrogen for storage at −80°C until use.

Processing samples for Illumina sequencing

For each liver sample, total RNA was extracted using Qiazol (Qiagen) as per manufacturer's instructions and DNase treated with DNA-free (Ambion). Polyadenylated mRNA was enriched from the total RNA using the PolyATract mRNA isolation system (Promega). Directional double-stranded cDNA was generated according to the method of Parkhomchuk et al. (2009), using the Superscript Double-Stranded cDNA Synthesis kit (Invitrogen), with uracil substituted for thymine in the second strand; 250 ng of double-stranded cDNA was fragmented by sonication and a sequencing library prepared for the Illumina platform using the Paired End Oligo Only kit (Illumina) according to the manufacturer's instructions. Strand specificity was then introduced by digestion of the second strand of cDNA using uracil-N-glycosylase. Each library was PCR-amplified using Illumina's PE primers. Size selection was performed by 2% agarose gel electrophoresis and 200–300-bp fragments were extracted using a Minelute gel purification kit (Qiagen). Libraries were sequenced on an Illumina GAIIx in the Genomics Core facility of the Cambridge Research Institute.

Preprocessing and low-level analysis

Following sequencing, reads from each library were aligned to a reference transcriptome using Bowtie. For libraries generated from tissue samples taken from the F0 mice, the reference was either the C57BL/6J or CAST/EiJ reference transcriptome, as appropriate. For the F1 mice, we aligned reads to a reference that contained both the C57BL/6J and the CAST/EiJ transcriptomes. In all cases, MMSEQ was used to estimate gene expression levels and, in the case of the F1 samples, to estimate allele-specific gene expression levels. We normalized the expression estimates for the F0 data using the approach of Anders and Huber (2010).

While the mapping of reads to the individual transcriptomes is straightforward for the F0 mice, the mapping of reads for the F1 hybrids is more complicated since the C57BL/6J and the CAST/EiJ alleles of each transcript only differ in a small number of positions. To assess whether this was a problem, we considered reads generated from two F0 libraries (one C57BL/6J and one CAST/EiJ) and combined them to generate a simulated F1. We then compared, for each gene, the expression estimate for the C57BL/6J allele in the simulated F1 hybrid with the expression estimate of the same gene for the F0 sample (Supplemental Fig. 2). We observed a high correlation between the two measurements. Furthermore, when we looked at the ratio of differential expression in the F0 mice and compared it with the ratio of allele-specific expression in the F1 mice, there was again good concordance (Supplemental Fig. 2). Both of these observations provide confidence in our expression estimation strategy.

To assess the quality of the data generated, we calculated the Spearman correlation between the gene expression levels across all 24 lanes of data sequenced before performing Principal Components Analysis (PCA). As can be seen in Supplemental Fig. 1, the samples clustered by strain as expected. Information about the total number of reads obtained in each lane, the proportion of reads mapped to the transcriptome, and the gene expression counts can be found in Supplemental Tables 11 and 12.

Finally, we defined genes as expressed using the following criteria. A gene is defined as expressed in the F0 mice if the expression estimates of all samples are >0 or if the expression estimate in at least one of the 12 F0 mice is ≥10. A gene is defined as expressed in both the F0 and F1 mice if one of the F0 criteria is satisfied and, additionally, the estimate is ≥10 across both alleles of the gene in at least one of the 6 F1i and in one of the 6 F1r mice.

Pyrosequencing

Genes were randomly selected for validation, following exclusion of genes that showed evidence of imprinting and those that showed highly variable expression levels between biological replicates. Single nucleotide variants (SNVs) were identified, and forward, reverse, and sequencing primers were designed to target each SNV using PyroMark Assay Design software (Qiagen). Each set of primers was tested for specificity in silico using BLAT and in the laboratory using quantitative PCR and using pyrosequencing on BL6xCAST genomic DNA. Primer sequences are given in Supplemental Table 13.

Pyrosequencing was performed on the Pyromark Q96 MD system (Qiagen), using the allele-specific quantification program. First, total RNA was isolated from the liver of one initial F1 cross and one reciprocal F1 cross mouse, and double-stranded cDNA was generated using the Superscript II double-stranded cDNA kit (Invitrogen). Three technical replicate PCR reactions were performed on each cDNA sample using one biotinylated and one nonbiotinylated primer. PCR products were purified and enriched using streptavidin sepharose beads on the Pyromark vacuum prep workstation (Qiagen). Pyrosequencing was performed on the enriched PCR products using Pyromark Gold Q96 reagents and Pyromark MD software (Qiagen). The Pyromark software determined an allelic ratio for each of three technical replicates, and the average was determined from all six replicates of each genomic locus.

Identification of highly variable genes

We identified genes with high variability within strains as follows. First we estimated the dispersion parameter under the negative binomial model described by Anders and Huber separately for each gene and strain (Anders and Huber 2010). As there is a dependency between dispersion and expression, we normalized the estimated dispersions by subtracting a first-order polynomial fitted through the scatterplot of the estimated dispersions plotted against the mean expression levels (Supplemental Fig. 4). We then ranked the expressed genes by their normalized dispersion levels and called the top 10% as highly variable (Supplemental Table 4).

Identifying imprinted genes

Only the hybrid (F1i and F1r) data were used to identify imprinted genes. Briefly, for each gene we introduce the following notation:

n_j^{in} = expression summed across both alleles for the j^{th} F1 initial cross replicate
z_j^{in} = expression of the C57BL/6J allele for the j^{th} F1 initial cross replicate
n_j^{re} = expression summed across both alleles for the j^{th} F1 reciprocal cross replicate
z_j^{re} = expression of the C57BL/6J allele for the j^{th} F1 reciprocal cross replicate

where $j = 1,\ldots,6$.

Subsequently, we assume that each count follows a beta-binomial distribution:

$$z_j^{in} \sim Bi\left(n_j^{in}, p_j^{in}\right) \text{ where } p_j^{in} \sim Be(a_1, b_1), \text{ and}$$
$$z_j^{re} \sim Bi\left(n_j^{re}, p_j^{re}\right) \text{ where } p_j^{re} \sim Be(a_2, b_2).$$

We can then model the null (H_0) and alternative (H_1) hypotheses of no imprinting and imprinting, respectively, using the following parameterizations:

$$H_0 : a_1 = a_2,\ b_1 = b_2 \text{ and } H_1 : b_1 = a_2,\ b_2 = a_1.$$

To discriminate between the two hypotheses, we estimated the parameters using a maximum likelihood based approach, calculated the maximum likelihood at these values, and calculated the likelihood ratio between them.

Since the null and alternative models are not nested we cannot compare the ratio to the quantiles of a χ^2 distribution. Instead we determined the distribution of the likelihood ratios under the null hypothesis of no imprinting. To do this, for each gene, we used data from the initial cross to calculate the corresponding maximum likelihood estimates for a_1 and b_1 and then simulated data from a reciprocal cross drawn from the distribution with these parameters. Using the real initial cross and simulated reciprocal cross data we then calculated the likelihood ratio using the procedure described above. We took the distribution of likelihood ratios obtained using this approach as our null model under the hypothesis of no imprinting (Supplemental Fig. 9). Given this, we assigned a *P*-value to each gene in the following way:

- *P*-value of 0 if the likelihood ratio is above the highest value of the null
- *P*-value of 1/n if between the first and the second highest values of the null, 2/n if between the second and third. . .

We corrected the *P*-values for multiple testing using the Benjamini-Hochberg procedure and adjudged that a gene showed evidence for being imprinted if the corrected *P*-value was <0.05.

Finally, to assess the quality of our method, we considered how well we could identify parent-of-origin effects for genes on the X chromosome. Here, we expect that all genes should be expressed from the maternal allele. Indeed, we find that, of the 284 genes expressed on the X chromosome, 268 (94%) showed the expected pattern, providing confidence in our analysis and our mapping strategy.

Classifying steady-state expression levels

To classify gene expression levels into different regulatory categories, for each gene, we introduce the following notation:

x_i = expression of the gene in the i^{th} C57BL/6J F0 mouse
y_i = expression of the gene in the i^{th} CAST/EiJ F0 mouse
n_j = number of reads mapping across both alleles in the j^{th} F1 hybrid
z_j = number of reads mapping to the C57BL/6J allele in the j^{th} F1 hybrid

Here i takes values between 1 and 6, and j takes values between 1 and 12, since we have pooled the initial and reciprocal crosses together (after removing imprinted genes). Subsequently, we make the following distributional assumptions:

$$x_i \sim Po(\mu_i),\ y_i \sim Po(\nu_i), \text{ and } z_j \sim Bi(n_j, p_j).$$

Further, we impose the following prior distributions upon μ, ν, and p:

$$\mu_i \sim Ga\left(r, \frac{p_\mu}{1-p_\mu}\right),\ \nu_i \sim Ga\left(r, \frac{p_\nu}{1-p_\nu}\right), \text{ and } p_j \sim Be(\alpha, \beta).$$

The marginal distributions of x_i, and y_i, are negative binomial and the marginal distribution of z_j is beta-binomial. Additionally, we note that r reflects the over-dispersion (relative to a Poisson distribution); this parameter is estimated a priori using the approach of Anders and Huber. Subsequently, different constraints upon the parameters can be imposed to describe the following biological situations:

$$\text{Conserved: } p_\mu = p_\nu \text{ and } \alpha = \beta,$$

$$\text{Cis: } p_\mu \neq p_\nu \text{ and } \frac{\alpha}{\alpha+\beta} = \frac{\frac{p_\mu}{1-p_\mu}}{\frac{p_\mu}{1-p_\mu} + \frac{p_\nu}{1-p_\nu}},$$

$$\text{Trans: } p_\mu \neq p_\nu \text{ and } \alpha = \beta,$$

$$\text{Cis \& Trans: } p_\mu \neq p_\nu \text{ and } \alpha \neq \beta.$$

We allocated each gene into one of these four categories. To do so, for each gene we fitted the four models described above to the data by maximizing the likelihood function. After doing this, we used the Bayesian Information Criterion (BIC) to determine which of the four models best fitted the data for each gene.

Subdivision of the *cis* + *trans* category

We subdivided genes allocated to the *cis* and *trans* class into four subclasses using the following approach. For the i^{th} gene we defined x_i as the average $\log_2$ fold change for the F0 data and y_i as the average $\log_2$ allelic ratio for the F1 data. Subsequently, we used the following criteria:

1) opposite—*cis* stronger: $(0 < y_i < x_i)$ OR $(0 > y_i > x_i)$
2) opposite—*trans* stronger: $(x_i < 0 < y_i)$ OR $(y_i < 0 < x_i)$
3) same—*cis* stronger: $(0 < x_i < y_i < 2x_i)$ OR $(0 > x_i > y_i > 2x_i)$
4) same—*trans* stronger: $(0 < 2x_i < y_i)$ OR $(0 > 2x_i > y_i)$.

Data access

The RNA-seq data generated in this study have been submitted to the EBI Array Express repository with accession number E-MTAB-1091. Processed count data can be found in Supplemental Tables.

Acknowledgments

We thank Stephen Watt for help in preparing the libraries prior to sequencing, the CRUK CRI Genomics and Biological Resources Facilities, and David Adams (Wellcome Trust Sanger Institute) for prior access to the genome and transcriptome sequences for the mouse strains used in this study. We also acknowledge helpful comments and suggestions from Simon Anders, Wolfgang Huber, Jenny Tung, Luis Barreiro, Athma Pai, Yoav Gilad, and members of the Marioni, Odom, Flicek, and Brazma groups. This research was supported by the European Research Council, EMBO Young Investigator Program, Hutchinson Whampoa (D.T.O.), Cancer Research UK (S.L.B., K.S., E.T., D.T.O.), University of Cambridge (S.L.B., K.S., E.T., D.T.O.), the Wellcome Trust (WT095908 and WT098051; D.T., P.F., D.T.O.), the FP7 HEALTH grant from the European Commission GEUVADIS (grant agreement 261123) (A.G., A.B.), and EMBL (A.G., D.T., A.B., P.F., J.C.M.).

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11:** R106. doi: 10.1186/gb-2010-11-10-r106.
Babak T, Deveale B, Armour C, Raymond C, Cleary MA, van der Kooy D, Johnson JM, Lim LP. 2008. Global survey of genomic imprinting by transcriptome sequencing. *Curr Biol* **18:** 1735–1741.
Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof H-P. 2007. GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res* **35:** W186–W192.
Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome Res* **20:** 180–189.
Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.
Cooper GM, Stone EA, Asimenos G; NISC Comparative Sequencing Program, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constrain in mammalian genomic sequence. *Genome Res* **15:** 901–913.
DeVeale B, van der Kooy D, Babk T. 2012. Critical evaluation of imprinted gene expression by RNA-Seq: A new perspective. *PLoS Genet* **8:** e1002600. doi: 10.1371/journal.pgen.1002600.
Doss S, Schadt EE, Drake TA, Lusis AJ. 2005. *Cis*-acting expression quantitative trait loci in mice. *Genome Res* **15:** 681–691.
Emerson JJ, Hsieh LC, Sung HM, Wang TY, Huang CJ, Lu HH, Lu MY, Wu SH, Li WH. 2010. Natural selection on *cis* and *trans* regulation in yeasts. *Genome Res* **20:** 826–836.
Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. *Nucleic Acids Res* **40:** D84–D90.
Gibson G, Weir B. 2005. The quantitative genetics of transcription. *Trends Genet* **21:** 616–623.
Gordon KL, Ruvinsky I. 2012. Tempo and mode in evolution of transcriptional regulation. *PLoS Genet* **8:** e1002432. doi: 10.1371/journal.pgen.1002432.
Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C. 2010. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329:** 643–648.
Haig D. 2004. Genomic imprinting and kinship: How good is the evidence? *Annu Rev Genet* **38:** 553–585.
Landry CR, Wittkopp PJ, Taubes CH, Ranz JM, Clark AG, Hartl DL. 2005. Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of *Drosophila*. *Genetics* **171:** 1813–1822.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.
Lees-Murdock DJ, Walsh CP. 2008. DNA methylation reprogramming in the germ line. *Adv Exp Med Biol* **626:** 1–15.
Lemos B, Araripe LO, Fontanillas P, Hartl DL. 2008. Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proc Natl Acad Sci* **105:** 14471–14476.
Li Y, Sasaki H. 2011. Genomic imprinting in mammals: Its life cycle, molecular mechanisms and reprogramming. *Cell Res* **21:** 466–473.
Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: From SNPs to phenotypes. *Trends Genet* **27:** 72–79.
McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp PJ. 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Res* **20:** 816–825.
Parkhomchuk D, Boradina T, Amstislavsky V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123. doi: 10.1093/nar/gkp596.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.
Schrem H, Klempnauer J, Borlak J. 2002. Liver-enriched transcription factors in liver function and development. Part I: The hepatocyte nuclear factor network and liver-specific gene expression. *Pharmacol Rev* **54:** 129–158.
Stern DL, Orgogozo V. 2008. The loci of evolution: How predictable is genetic evolution? *Evolution* **62:** 2155–2177.
Tirosh I, Reikhav S, Levy AA, Barkai N. 2009. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324:** 659–662.
Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Sliverman JS, Powell J, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39:** 31–40.
Turro E, Su SY, Goncalves A, Coin LJ, Richardson S, Lewin A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12:** R13. doi: 10.1186/gb-2011-12-2-r13.
Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19:** 327–335.
Wang X, Soloway PD, Clark AG. 2011. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics* **189:** 109–122.
Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in *cis* and *trans* gene regulation. *Nature* **430:** 85–88.
Wittkopp PJ, Stewart EE, Arnold LL, Neidhert AH, Haerum BK, Thompson EM, Akhras S, Smith-Winberry G, Shefner L. 2009. Intraspecific polymorphism to interspecific divergence: Genetics of pigmentation in *Drosophila*. *Science* **326:** 540–544.
Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* **8:** 206–216.
Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148:** 816–831.

Received April 25, 2012; accepted in revised form August 15, 2012.