nature genetics

Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance

James J Crowley[1,10], Vasyl Zhabotynsky[1,10], Wei Sun[1,2,10], Shunping Huang[3], Isa Kemal Pakatci[3], Yunjung Kim[1], Jeremy R Wang[3], Andrew P Morgan[1,4,5], John D Calaway[1,4,5], David L Aylor[1,9], Zaining Yun[1], Timothy A Bell[1,4,5], Ryan J Buus[1,4,5], Mark E Calaway[1,4,5], John P Didion[1,4,5], Terry J Gooch[1,4,5], Stephanie D Hansen[1,4,5], Nashiya N Robinson[1,4,5], Ginger D Shaw[1,4,5], Jason S Spence[1], Corey R Quackenbush[1], Cordelia J Barrick[1], Randal J Nonneman[1], Kyungsu Kim[2], James Xenakis[2], Yuying Xie[1], William Valdar[1,4], Alan B Lenarcic[1], Wei Wang[3,9], Catherine E Welsh[3], Chen-Ping Fu[3], Zhaojun Zhang[3], James Holt[3], Zhishan Guo[3], David W Threadgill[6], Lisa M Tarantino[7], Darla R Miller[1,4,5], Fei Zou[2,11], Leonard McMillan[3,11], Patrick F Sullivan[1,5,7,8,11] & Fernando Pardo-Manuel de Villena[1,4,5,11]

Complex human traits are influenced by variation in regulatory DNA through mechanisms that are not fully understood. Because regulatory elements are conserved between humans and mice, a thorough annotation of *cis* regulatory variants in mice could aid in further characterizing these mechanisms. Here we provide a detailed portrait of mouse gene expression across multiple tissues in a three-way diallel. Greater than 80% of mouse genes have *cis* regulatory variation. Effects from these variants influence complex traits and usually extend to the human ortholog. Further, we estimate that at least one in every thousand SNPs creates a *cis* regulatory effect. We also observe two types of parent-of-origin effects, including classical imprinting and a new global allelic imbalance in expression favoring the paternal allele. We conclude that, as with humans, pervasive regulatory variation influences complex genetic traits in mice and provide a new resource toward understanding the genetic control of transcription in mammals.

The genetic basis of most phenotypic variation can be assigned to variation in protein-coding, RNA or regulatory sequences. The importance of regulatory sequence has become increasingly apparent in recent studies comparing divergent taxa and populations[1–4] and through the identification of thousands of SNPs that, although not predicted to change protein structure, are nonetheless strongly associated with human diseases and biomedical traits[5–8]. Here we investigated the effects of genetic variation and parental origin on gene expression in multiple tissues in laboratory mice. The study design maximized the level of genetic variation while concurrently enhancing the capacity to assign transcripts to either one of the two parental alleles. Examination of allele-specific expression (ASE) can be used to detect allelic imbalance in transcription in heterozygous mice, a process that requires genetic or epigenetic variation in *cis*. Therefore, we designed our experiment to include reciprocal F_1 hybrids to detect and quantify statistically significant allelic imbalance in expression for as many genes as possible.

Previous publications have examined allelic imbalance in F_1 mice using RNA sequencing (RNA-seq) (**Supplementary Table 1**). Four studies examined brain[9–12], one reported multiple tissues[4], two used fetal placenta[13,14], one used adult liver[15] and one used whole embryo[16]. However, some of the conclusions of these RNA-seq studies have been controversial[17]. A particularly controversial issue is the number of mouse genes subject to imprinting. Previous consensus estimates placed the number of imprinted genes in mouse at 100–200 (ref. 18). An early application of RNA-seq in brain tissue yielded a small number of new imprinted transcripts[9], but 2 subsequent studies claimed identification of >1,300 new imprinted loci[10,11], including 347 autosomal genes with sex-specific imprinting[11]. A reanalysis did not replicate these claims[12].

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [3]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [5]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [6]Department of Molecular and Cellular Medicine, Texas A&M Health Science Center, College Station, Texas, USA. [7]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA. [8]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [9]Present addresses: Department of Biological Sciences, North Carolina State University, Raleigh, North Carolina, USA (D.L.A.) and Department of Computer Science, University of California, Los Angeles, Los Angeles, California, USA (W.W.). [10]These authors contributed equally to this work. [11]These authors jointly supervised this work. Correspondence should be addressed to F.P.-M.d.V. (fernando@med.unc.edu).

Received 17 September 2014; accepted 26 January 2015; published online 2 March 2015; corrected after print 16 April 2015; doi:10.1038/ng.3222

In the context of these findings, we sought to improve knowledge of the control of gene expression in mouse. To maximize generalizability, we studied related but divergent genomes. We selected three inbred mouse strains (CAST/EiJ, PWK/PhJ and WSB/EiJ) representative of three subspecies within the *Mus musculus* species group (*M. m. castaneus*, *M. m. musculus* and *M. m. domesticus*, respectively). We chose these strains to maximize the level of genetic diversity (for example, 27.7 million SNPs and 4.6 million indels vary in these strains[4]), the number of genes with expressed SNPs and/or indels (31,259 of 36,817 Ensembl v37 genes) and the number of such variants per gene (mean of 19.9, s.d. of 26.9).

We conducted all possible pairwise crosses to form a 3 × 3 diallel (**Fig. 1**) and measured gene expression in brain, liver, kidney and lung with age- and sex-matched biological replicates for each of the nine possible genotypic combinations. We used RNA-seq to measure ASE in brain and microarrays to assess gene expression in brain, liver, kidney and lung. Inclusion of the array data allowed a detailed comparison of two major platforms for expression analysis, determination of the proportion of genetic effects that are missed by examining a single tissue and estimation of the degree to which strain, sex and parent-of-origin effects in brain are reproduced in other tissues.

In designing this experiment, we attempted to optimize the discovery of regulatory variation and to address potential pitfalls (**Supplementary Table 2**). In particular, we included three genomes instead of two, allowing us to generalize our conclusions, to estimate the proportion of variants that have a *cis* regulatory effect and to assist the aims of large-scale projects such as the International Knockout Mouse Consortium[19], Collaborative Cross[20] and Diversity Outbred[21]. We also increased the depth of sequencing and the number of replicates and included both sexes to improve power to detect ASE. We developed a new approach to diploid genome alignment to customized genomes ('pseudogenomes')[22–24] created from the highest quality and most current genomic data available[4].

Allelic imbalance in expression for an F_1 mouse requires the presence of a genetic or epigenetic regulatory variant acting in *cis*, as *trans*-acting factors have an equal opportunity to affect both alleles (**Supplementary Fig. 1**). Regulatory variation in *cis* causes differential expression from the linked allele, which is detected by a statistically significant imbalance in the ASE derived from each parental allele in an F_1 mouse (**Supplementary Fig. 2**). We observe *cis* regulatory effects for >80% of all testable genes. We also found that the number of imprinted genes was not substantially different from historical estimates, but we report a new genome-wide parent-of-origin allelic imbalance favoring expression of the paternal allele.

RESULTS

Major drivers of differential gene expression in mice

We hybridized brain, liver, kidney and lung RNA samples from the same mice used for RNA-seq to expression microarrays. Clustering of gene expression data from 384 microarrays (4 tissues × 96 samples) partitioned the samples perfectly by tissue (**Supplementary Fig. 3a**), indicating that the predominant predictor of gene expression is tissue type, even in the presence of extreme genetic diversity and representation of both sexes. After tissue, the samples partitioned by strain, then by parent of origin and finally by sex. Microarray data also showed that, across different tissues, strain effects are commonly shared (**Supplementary Fig. 3b**), suggesting that regulatory variation across diverse tissues often acts in a similar manner. Brain RNA-seq total read counts and microarray intensity values were highly correlated (median $r = 0.86$, range of 0.84–0.87).

Within each tissue, the overwhelming driver of differential gene expression was strain; this effect greatly exceeded the effects from parent of origin and sex (**Fig. 2**). For RNA-seq, the first two principal components accounted for ~30% of the total variation in autosomal total read count (TReC). The remaining top ten principal components were also strongly determined by strain and, to a far lesser extent, parent of origin and sex, with no notable effects from the barcodes used for multiplexing (**Supplementary Table 3**).

Within each tissue, the three inbred strains formed an equilateral triangle with the F_1 samples located midway between the corresponding parental strains (**Fig. 2**). This indicates that there was no overall bias in the alignment of RNA-seq reads to these three equally divergent genomes. We also determined that the genetic architecture of regulatory variation in laboratory mice was mostly additive, as the F_1 samples would not be located midway between the parental strains if dominance and parent-of-origin effects predominated.

Cis regulatory variation is pervasive in diverse mice

We found *cis* regulatory effects for 11,287 autosomal genes (89% of testable genes). More than 75% of these genes showed consistent additive effects, defined by having an additive TReC effect and an additive allele-specific read count (ASReC) effect in the same direction within a cross. For example, *Mad1l1* showed allelic imbalance in expression for all three crosses, indicating that, at the *cis* level, the PWK/PhJ allele is stronger than the WSB/EiJ allele, which in turn is stronger than the CAST/EiJ allele (**Supplementary Fig. 4**). Furthermore, this *cis* effect is consistent with the differential gene expression of the parental inbreds, and the level of gene expression in the F_1 mice can be explained as an additive effect. Some fraction of *cis* regulatory variants create strain effects that are undetectable in TReC or inconsistent between TReC and ASReC, owing to dominance and other effects. For example, *Fos* showed allelic imbalance in all F_1 mice in a manner consistent with TReC in the parental inbreds, but the total levels of gene expression in the F_1 mice were best explained as an effect of dominance or overdominance (**Supplementary Fig. 5**). Copy number variation can also lead to inconsistency between TReC and ASReC and result in underestimation of the number of genes with *cis* effects.

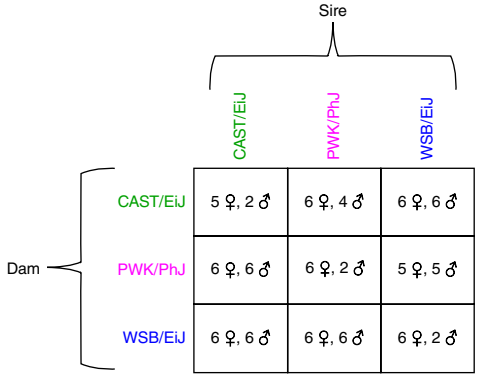

Figure 1 Diallel crossing scheme and sample sizes. We selected three divergent inbred strains representative of three subspecies within the *M. musculus* species group. We generated offspring from all possible pairwise crosses to form a 3 × 3 diallel, including age- and sex-matched biological replicates for each of the nine possible genotypic combinations. Mice were aged to 23 d and killed, and total RNA was extracted from whole brain, liver, kidney and lung. The sample size shown is for RNA-seq (52 females, 39 males). RNA-seq was performed on RNA extracted from brain, and microarrays were run on RNA extracted from brain, liver, kidney and lung.

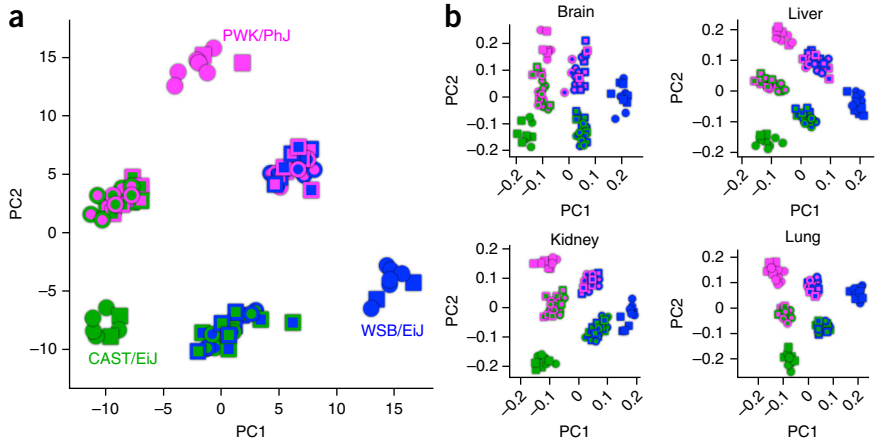

Figure 2 Principal components (PCs) of brain RNA-seq and microarray expression levels across four tissues. Each point represents one mouse, with shape indicating sex (circle, female; square, male) and color indicating genotype. For the F_1 mice, the outer color indicates the maternal strain and the inner color indicates the paternal strain. (**a**) PC1 versus PC2 of the brain RNA-seq TReC for all autosomal genes. The three inbred strains form a near-perfect triangle with the F_1 samples located between their corresponding parental strains. PC1 and PC2 account for 31% of the variance in TReC, indicating that genetic background is the overwhelming driver of gene expression difference, with its effect greatly exceeding those of parent of origin and sex. (**b**) PC1 versus PC2 of microarray expression values for all autosomal genes across four tissues. The pattern seen in brain extends to multiple diverse tissues.

Of the 11,287 autosomal genes with *cis* regulatory effects, 4,113 (36%) were detected for all 3 pairs of strains, 5,065 (45%) were detected for 2 pairs and 2,109 (19%) were detected for 1 pair (**Fig. 3a**). Notably, all three subspecies contributed similarly to differential gene expression, indicating that there was no overall bias in read alignment to any one genome. Furthermore, the fold-change distribution of allelic imbalance effect sizes showed a similar pattern among the three crosses, and there was minimal skewing in the ratio of upregulated to downregulated genes in any given cross (**Fig. 3b**). We saw a similar pattern with the microarray data across the four tissues analyzed (**Supplementary Fig. 6**).

Phenotypic consequences and human relevance

To test the potential consequences of *cis* regulatory variation, we compared our results to a comprehensive set of knockout mouse phenotypes for 6,039 different genes and 29 phenotype dimensions (see URLs). Brain-expressed genes with *cis* regulatory effects were significantly more likely to be associated with a behavioral or neurological phenotype in knockout mice ($P = 0.012$) than brain-expressed genes with no *cis* effect. Furthermore, we found no such enrichment for the 1,348 genes that result in no overt aberrant phenotype after being knocked out ($P = 0.56$) or those associated with the 27 other phenotype dimensions.

To test the human relevance of mouse *cis* regulatory variation, we compared our results to those for human expression quantitative trait locus (eQTL) studies. These comparisons were restricted to only the genes that have a one-to-one ortholog for mouse and human ($n = 15{,}312$ genes; see URLs). Brain-expressed genes with a *cis* regulatory effect in mouse were much more likely to have a human peripheral blood eQTL ($P = 7.8 \times 10^{-10}$)[25]. Published human brain eQTL studies had much smaller samples sizes; nonetheless, when comparing our results to a meta-analysis[26] of 5 available data sets (total $n = 439$), we observed consistent enrichment ($P = 0.04$).

Proportion of SNPs with *cis* regulatory effects

In contrast to previous F_1 RNA-seq studies, we included three genomes in our experimental design to allow multiple pairwise comparisons. In our experiment, for >90% of the genome, pairwise comparisons were possible between the different subspecies (*M. m. domesticus*, *musculus* or *castaneus*), whereas, for the remainder of the genome, just one subspecies was represented (*M. m. domesticus* or *musculus*)[27]. Therefore, we could make six comparisons: three between genomic regions of different subspecific origin and three between regions of the same subspecific origin. For each comparison, we examined the relationship between sequence diversity (SNPs/kb) and the fraction of genes that showed differential gene expression (additive, consistent strain effects). The result was a positive logarithmic correlation (**Fig. 4**), indicating that the number of functional regulatory variants per kilobase increases as the number of total variants per kilobase increases. Furthermore, within each pairwise comparison, sequence diversity was correlated with the fraction and magnitude of genes with differential gene expression (**Supplementary Fig. 7**), and this correlation replicated in all four tissues.

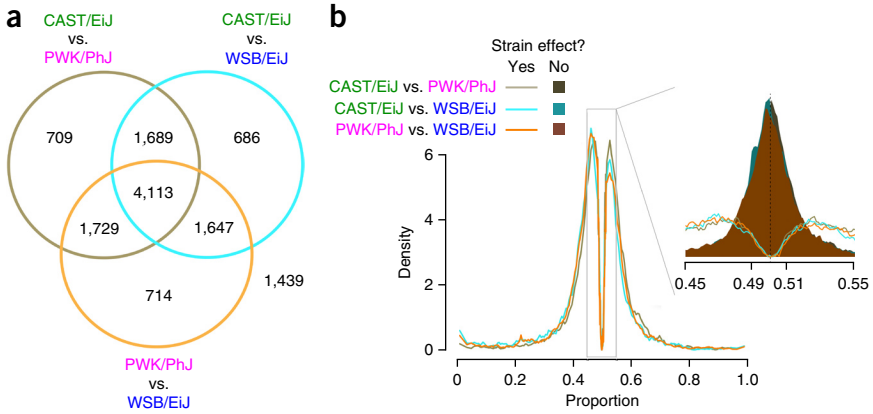

Figure 3 Balanced contribution of different subspecies to the identification of *cis*-regulated genes. (**a**) Venn diagram showing the number of genes with allelic imbalance (false discovery rate (FDR) < 0.05) in each cross and the relationship to other crosses. (**b**) Distribution of allelic imbalance effect sizes for the 11,287 autosomal genes that showed allelic imbalance in expression for at least one cross. In each cross, the proportion is the fraction of allele-specific reads from the strain listed second in the legend (i.e., PWK/PhJ or WSB/EiJ). The inset magnifies the distribution of effect sizes in the vicinity of 0.5 and provides, in the background, the distribution of effect sizes for genes that did not reach statistical significance for a strain effect (filled distributions).

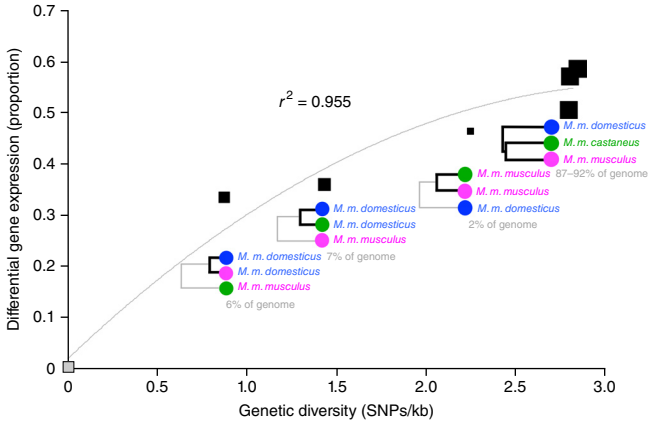

Figure 4 Differential gene expression is positively correlated with sequence diversity at multiple evolutionary scales. Each square indicates the relationship between the local level of sequence diversity (SNPs/kb) and the fraction of genes that show differential gene expression (proportion of genes with additive, consistent strain effects), for regions of the genome with the same or different subspecific origin (indicated by dendrograms). Colored circles represent strain (magenta, PWK/PhJ; blue, WSB/EiJ; green, CAST/EiJ), and colored text represents the subspecific origin in the regions of the genome considered (magenta, *M. m. musculus*; blue, *M. m. domesticus*; green, *M. m. castaneus*). For each of the six pairwise comparisons, only expressed genes with allele-specific information were considered and only SNPs within the entire gene body (±10 kb) were included. The portion of the genome considered for each of these six comparisons was approximately, from left to right, 50 Mb, 150 Mb, 175 Mb and 2.25 Gb for the final three comparisons.

Each *cis* eQTL identified in this study was explained by at least one regulatory variant. Therefore, we could estimate the lower bound of the proportion of mutations that create a *cis* regulatory effect by dividing the number of *cis* eQTLs by the number of SNPs within genomic regions spanning all testable genes for a particular cross (**Supplementary Fig. 8**). The overall ratio was 0.10% (±0.02%), such that approximately 1 in 1,000 SNPs creates a *cis* regulatory effect. This estimate was stable across all crosses examined and across all regions independently of their phylogenetic origin. This estimate also generalized to genes of varying size and levels of expression.

Classical imprinting is incomplete and under genetic control

We identified 95 genes with significant evidence of imprinting (**Fig. 5a**; full gene list in the **Supplementary Data Set**). Significance was defined by a parent-of-origin effect q value <0.01 or a P value <0.01 combined with evidence of imprinting from a hidden Markov model (see Zou *et al.*[24] and the **Supplementary Note**). Imprinted genes were found on 16 chromosomes, with 62 of the 95 genes residing in well-known imprinting clusters (**Supplementary Fig. 9**). There were 52 new imprinted genes and 43 genes with previous evidence of imprinting (see URLs). Of 128 genes with previous evidence of imprinting in the literature, 73 could be evaluated

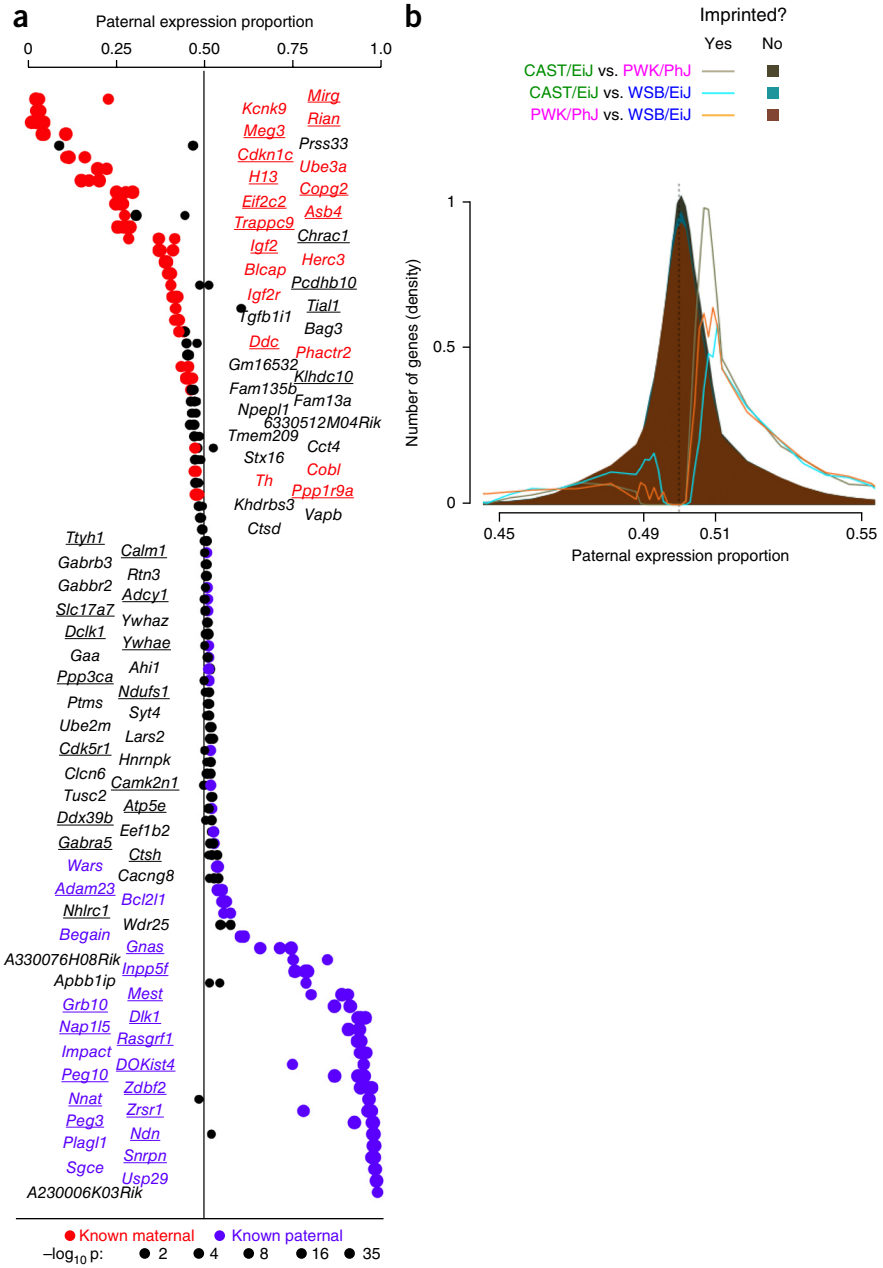

Figure 5 Imprinted genes in mouse brain. (**a**) Paternal expression ratio for 95 genes with a significant parent-of-origin effect. Each dot corresponds to a reciprocal cross (for example, CAST/EiJ × PWK/PhJ versus PWK/PhJ × CAST/EiJ), and dot size is proportional to the parent-of-origin effect P value. Genes known from the literature to be maternally expressed are shown in red, those known to be paternally expressed are shown in blue and new imprinted genes are shown in black (n = 54 new genes). Genes with a strain × parent of origin effect are underlined (n = 47 genes). (**b**) Distribution of the parental expression proportion in the vicinity of 0.5 for genes that are imprinted (lines) and, in the background, genes that did not reach statistical significance for parent of origin–dependent expression (filled distributions).

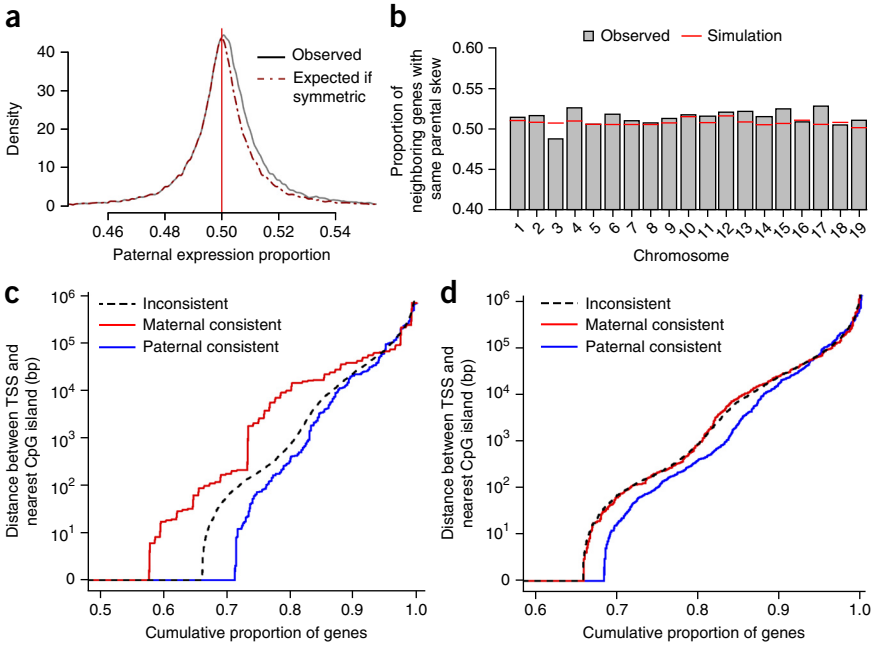

Figure 6 Global allelic imbalance in favor of the paternal allele. (**a**) Distribution of the proportion of paternal expression for all genes, except the 95 imprinted genes described in **Figure 5**. The distribution reflects aggregate data for ~10,000 genes × 3 crosses × 2 sexes. The dashed red line represents a reflection of the values to the left of 0.5 (the expectation if no paternal skew were present). (**b**) Genes with consistent allelic imbalance (found in all three crosses) are clustered on most autosomes. The red line denotes the expected proportion of clustering based on the number of genes with consistent paternal or maternal overexpression on every autosome. (**c**) Genes with consistent paternal overexpression in all 3 crosses and both sexes (n = 467) tend to be closer to CpG islands, whereas those with consistent maternal overexpression (n = 116) tend to be farther away, relative to inconsistent genes (n = 9,540). Plotted is the cumulative proportion of genes with a given distance between the TSS and the nearest CpG island. (**d**) Expanded analysis including genes not fully consistent in both sexes but still consistent in all three crosses. Genes with consistent paternal overexpression (n = 3,338) retain enrichment for CpG islands, whereas those with consistent maternal overexpression (n = 1,631) are not different from inconsistent genes (n = 5,154).

(expressed and containing exonic variation) and 42 (58%) were identified as being imprinted. The remaining 31 genes were sufficiently expressed (median TReC = 809, median ASReC = 143) for evaluation but did not meet the criteria for parent of origin–dependent expression (median P = 0.37, range = 0.01–0.97), suggesting tissue-specific imprinting, lack of imprinting in brain or strain effects on imprinting.

Allele-specific RNA-seq data allowed quantification of the strength of imprinting for each gene. For most genes, imprinting was incomplete. For maternally expressed genes, maternal reads represented an average 67% of the ASReC (range of 51.5–97.9%). For paternally expressed genes, paternal reads represented an average 75.6% of the ASReC (range of 50.6–99.7%). The strength of imprinting was highly replicable, with a mean variance of 3.2% within a cross. Of the 95 imprinted genes, 47 showed a strain effect modifying the strength of imprinting (strain × parent of origin effect). We divided these 47 genes into 2 classes: those for which the differential gene expression could be explained by a single strain effect (n = 11) and those for which it could not, suggesting the existence of a more complex model (n = 36) (**Supplementary Table 4**).

Global allelic imbalance in favor of the paternal allele

Imprinted genes were 1.5 times more likely to be expressed from the paternal than the maternal allele (**Fig. 5b**). This finding is consistent with the observation that paternal expression predominates in brain whereas maternal expression predominates in placenta[9]. To test whether this asymmetry in parent-of-origin effects extended beyond imprinted genes, we estimated the parent-of-origin effect for each cross and each sex separately. We found that 54–60% of genes showed higher expression from the paternal allele, a proportion significantly different from the expectation of 50% ($P = 5.9 \times 10^{-24}$; **Fig. 6a** and **Supplementary Table 5**). We also observed that genes with higher expression from one parental allele tended to cluster (**Fig. 6b**). Among the 19 autosomes, 15 had a higher proportion of genes whose neighbor had the same parental skew in expression than expected by chance ($P = 9.6 \times 10^{-3}$, binomial test).

We calculated a rough estimate of the number of genes with paternal overexpression simply by taking the difference between the number of genes with higher paternal expression and the number of genes with higher maternal expression. For example, for female CAST/EiJ × PWK/PhJ reciprocal hybrids, there were 1,652 more genes with allelic imbalance in favor of the paternal allele (6,790 paternally overexpressed genes minus 5,138 maternally overexpressed genes). The excess of genes with paternal overexpression ranged between 938 and 2,500 (across reciprocal crosses stratified by sex; **Supplementary Table 5**). However, these numbers likely represent an underestimation because we conservatively assumed that all genes with higher maternal expression occurred by chance, while some proportion are not due to chance. Although we can estimate the number of genes with paternal overexpression, we lack sufficient power to identify all genes with modest parental overexpression while correcting for multiple testing.

To identify genomic features associated with parentally overexpressed genes, we first selected genes with consistent paternal or maternal overexpression in the three reciprocal crosses (with or without stratification by sex). These genes were not significantly clustered with known imprinted genes. However, when we examined the proximity of these genes to CpG islands, we found that the transcription start sites (TSSs) of genes with consistent overexpression of the paternal allele in all 3 crosses (n = 467 with and 3,338 without stratification by sex) were closer to CpG islands ($P < 1 \times 10^{-5}$) than the TSSs for the remaining genes (**Fig. 6c**,**d**). We did not observe this effect among genes with consistent maternal overexpression (n = 116 and 1,631; P = 0.60). Note that, for the more restrictive group (consistently expressed in both sexes within each cross), there was further enrichment for genes with a TSS near a CpG island among paternally overexpressed genes and a significant depletion of genes with a TSS near a CpG island among maternally overexpressed genes ($P = 1 \times 10^{-5}$; **Fig. 6c**).

For genes consistently overexpressing the paternal allele, we observed that the size of the strain effect was significantly smaller than for other genes ($P < 1.2 \times 10^{-4}$), implying that *cis*-acting regulatory elements have less impact on these genes. Interestingly, the proximity of a CpG island to a TSS was associated with smaller additive strain effect sizes, and genes with a TSS that overlapped a CpG island were also clustered in the genome. We conclude that, in addition to the statistically significant allelic imbalance observed at the gene level (imprinting), there is an association between the proximity of a CpG island to a TSS and a pervasive allelic imbalance favoring expression of the paternal allele in brain; this suggests that parent of origin–dependent methylation may be implicated in this phenomenon.

We were able to support this claim using a recently published whole-genome parent-of-origin brain DNA methylation data set from reciprocal hybrids of 129X1/SvJ and CAST/EiJ mice[28]. Genes with consistent overexpression from the paternal allele were closer to CpG islands that were preferentially methylated on the maternal allele (**Supplementary Fig. 10**). This observed relationship between paternal overexpression and nearby maternal methylation is not simply the result of inherent differences between CpG islands with a paternal versus maternal methylation bias[28].

Two forms of dosage compensation on the X chromosome

Gene expression on the X chromosome in mammals is believed to be subject to two forms of dosage compensation. The first equalizes the expression of X-linked genes in females and males[29,30], and the second equalizes the average expression of X-linked genes with the expression of autosomal genes[31]. In our data set, the overall level of X-chromosome gene expression was equivalent in males and females in all four tissues examined (**Supplementary Fig. 11a**). These data indicate that the silencing of one X chromosome in females equalizes the average expression of X-linked genes between females and males[29,30]. In addition, X-chromosome gene expression was equivalent to that for the autosomes in all four tissues examined (**Supplementary Fig. 11b**). These data support the hypothesis[31] that, during the evolution of mammalian sex chromosomes from a pair of autosomes, expression of X-linked genes was doubled to compensate for the degeneration of Y-chromosome homologs. We also observed an effect of genotype at *Xce* (X-chromosome–controlling element)[32] and a parent-of-origin effect in X-chromosome inactivation skewing in females (**Supplementary Fig. 12**)[33].

A total of 346 X-chromosome genes were found to possess a strain effect (77% of all expressed and testable genes), a rate slightly lower than that for autosomes. This difference is expected because of the reduction in power to detect effects on the X chromosome, as ASReC data can only be informative in female samples. Of the 527 testable X-linked genes, only 4 (0.76%) were differentially expressed between the sexes, a rate similar to the autosomes (0.28%). Overall, however, sex did account for ~12% of the variation in X-chromosome gene expression, with effects largely driven by one gene, *Xist*.

DISCUSSION

We find that more than 80% of mouse genes have expression levels dependent on genetic variation. The majority of these differentially expressed genes fit an additive model and are subject to regulatory variation acting in *cis*. These *cis* regulatory effects have functional consequences for mouse phenotypes and usually extend to the human ortholog. Furthermore, differential gene expression is positively correlated with sequence diversity at multiple evolutionary scales, and the proportion of mutations that create a *cis* regulatory effect has remained relatively constant as mouse subspecies evolved. Two types of parent-of-origin effects on gene expression were observed. First, we demonstrated that the number of classically imprinted genes is not substantially different from historical estimates. Second, we observed a global allelic imbalance in favor of expression of the paternal allele at a large number of genes associated with CpG islands. For most genes, imprinting is incomplete, and *cis*-acting mutations can modify the strength of imprinting. Furthermore, we conclude that regulation of gene expression on the X chromosome is similar to that for the autosomes and includes two forms of dosage compensation. Finally, we developed improved analytical tools with broad usefulness for RNA-seq analysis in many species (see URLs; **Supplementary Table 2**)[22–24]. These tools improve the power to detect allele-specific and parent-of-origin effects while minimizing false discoveries and reference bias, detect and correct spurious transcriptome inference due to RNA-seq read misalignment and allow analysis of expression on the X chromosome without chromosome-wide confounding effects. Finally, a new likelihood-based method to jointly analyze TReC and ASReC from inbred and F_1 mice (**Supplementary Fig. 2**) increases statistical power to detect genetic effects.

We found *cis* regulatory effects for 11,686 genes (85% of testable genes). This number exceeds all previous findings for mouse eQTL studies[34]. We found that the expression of most transcripts shows an additive pattern of inheritance, consistent with studies in mouse[35], human[36] and plant[37]. Interestingly, many genes have inconsistent patterns of inheritance between TReC and ASReC. We have determined that, when one of the strains used to create the reciprocal F_1 hybrid has a copy number gain, typically no SNPs and small indels are called in that strain in that genomic region[4]; this leads to allele-specific reads from that strain being undercounted. However, patterns of TReC—which are independent of variant calls—are still informative for copy number status.

Inbred mouse strains are assumed to possess a fixed genome across time, but mutations arise continuously. We observed two striking examples of *de novo* mutations altering gene expression via changes in gene dosage. Among the 96 samples included in the RNA-seq study, we identified one XO female caused by paternal nondisjunction (supported by genotyping) and another mouse with a ~250-kb duplication spanning 5 genes (**Supplementary Fig. 13**).

Pinpointing the genetic variants that underlie mouse QTLs has been challenging because the QTLs detected in experimental crosses often span hundreds of genes. The data described here can help investigators prioritize candidate genes on the basis of strain distribution patterns or tissue-specific expression. Furthermore, if differential expression of a particular gene is suspected to influence a phenotype, these data provide the means to create an 'allelic series', a set of animals bred intentionally to titrate the level of gene expression. This approach could complement or even incorporate gene-targeted knockout mice.

In humans, common disease-associated variants are enriched for regulatory DNA. Animal models for such regulatory variation are needed to provide a more detailed understanding of genotype-phenotype relationships. We have shown that eQTL patterns are often independent of species and tissue, such that *cis*-regulated genes in human blood often have a counterpart in the mouse ortholog, providing a tractable model to assess the effect of regulatory variation on phenotype.

We have provided a lower-bound estimate of the proportion of variants that have a *cis* regulatory effect. We estimate that at least 1 in every 1,000 SNPs creates a *cis* regulatory effect. Therefore, at least 47,000 regulatory variants are segregating in the Collaborative Cross[20] and Diversity Outbred[21] populations. These regulatory variants likely

contribute to the broad phenotypic distributions seen in those populations, and the small proportion of testable genes without regulatory variation (~15% in this study) are likely under selective pressure to maintain gene expression at a constant level. Furthermore, as human and mice average ~100 *de novo* mutations per generation[38,39], at least 1 in 10 offspring should have a new regulatory mutation. Given this proportion and the size of the human population, several million new regulatory variants are likely created each year.

There have been conflicting reports regarding the number of mouse genes subject to imprinting. If the definition of imprinting is restricted to genes that show significant allelic imbalance in expression favoring one parent, then our results indicate that the number of genes imprinted in mouse brain is in line with the historical consensus. However, parent-of-origin effects on gene expression appear to be asymmetric in mouse brain, with favored expression of the paternal allele. This affects many genes distributed across all the autosomes and is present in all three reciprocal crosses. The 467 genes that have consistent overexpression of the paternal allele in all 3 crosses and both sexes are strongly enriched for CpG islands near their TSSs and tend to show smaller strain effects relative to inconsistent genes (**Fig. 6**). In addition, genes with consistent overexpression of the paternal allele are associated with differentially methylated CpG islands (**Supplementary Fig. 10**). These observations suggest that differential parent of origin–dependent resetting of methylation marks during early development is likely the mechanism responsible for global allelic imbalance.

We hypothesize that this global imbalance is ancestral to classical imprinting. In other words, small differences in parental methylation at CpG islands close to the TSS may have been exploited by natural selection to create classical imprinting. We propose that the difference in size of strain effects between genes that are affected or not by this parent-of-origin effect could be explained by the fact that mutations in the promoters of genes of the later type are likely to create strong *cis* regulatory variants. On the other hand, mutations in CpG islands will only have an overall minor effect on overall methylation. Lastly, the global allelic imbalance in favor of expression of the paternal allele may partly explain why the majority of the newly identified imprinted genes described here (37 of 54) show modest overexpression of the paternal allele and may also explain the surprisingly large number of genes found in 2 previous controversial studies of imprinting[10,11].

We verified two forms of dosage compensation on the X chromosome. First, for most of the genes on the X chromosome, we found that males and females have similar expression. Although this has been demonstrated before using cell lines[40,41], here we provide additional evidence in primary tissue samples. Furthermore, we confirm that it is rare for genes to escape X inactivation in mouse, with this occurring for just 1.1% of the genes that could be tested, all of which have previously been identified[42–44]. This finding stands in sharp contrast to the scenario in human females, where ~15% of X-chromosome genes are biallelically expressed[45,46]. Second, we found that the overall level of X-chromosome expression is roughly equivalent to expression on the autosomes (Ohno's hypothesis)[31]. Ohno's hypothesis was initially supported by three microarray studies across several eutherian species[40,47,48] but then contradicted in 2010 by an RNA-seq analysis of mouse and human tissues[49], and this controversy remains, despite multiple recent studies[50–56]. The main factor contributing to disparate results across studies has been whether genes with low expression are considered[57,58]. Because genes with no or low expression in somatic tissues are more abundant on the X chromosome than on autosomes[50], their inclusion can lower median X:autosome expression ratios. Our analysis considers all genes on the X chromosome and clearly supports Ohno's hypothesis in mouse. This form of dosage compensation provides strong evidence that the level of gene expression is under evolutionary pressure.

In summary, our study demonstrates that in the laboratory mouse the vast majority of genes are subject to *cis* regulatory variation. Mouse models incorporating regulatory variation[20,21] should provide a powerful complement to null mutants[19] in the search for the mechanisms underlying complex genetic traits in humans.

URLs. Expression data can be viewed at http://csbio.unc.edu/gecco/. Scripts are provided to construct pseudogenomes (http://code.google.com/p/lapels/) and perform diploid alignment (http://code.google.com/p/suspenders/). An R package for jointly analyzing TReC and ASReC and to factor in X-inactivation skewing can be found at http://www.bios.unc.edu/~feizou/software/rxSeq. For detection and correction of spurious RNA-seq read misalignment (pseudogenes), access GeneScissors at http://csbio.unc.edu/genescissors/. Knockout mouse phenotypes were acquired from http://www.informatics.jax.org/phenotypes.shtml. Orthologous genes for human and mouse were identified from Ensembl (http://www.ensembl.org/info/genome/compara/homology_method.html) using the category "ortholog_one2one." Genes with previous evidence of imprinting were identified by creating a union of the databases from the following websites: http://www.geneimprint.com/, http://igc.otago.ac.nz/ and http://www.mousebook.org/catalog.php?catalog=imprinting.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Expression data can be acquired from the Gene Expression Omnibus (GEO) under accession GSE44555. RNA-seq data sets that passed quality control are available at the Sequence Read Archive (SRA) under accession SRP056236.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Mieczkowski, A. Brandt, E. Malc, M. Vernon, J. Brennan and M. Calabrese for helpful discussions. Major funding was provided by National Institute of Mental Health/National Human Genome Research Institute Center of Excellence for Genome Sciences grants (P50MH090338 and P50HG006582, co-principal investigators F.P.-M.d.V. and P.F.S.). This work was also supported by grants R01GM074175 (principal investigator F.Z.) from the National Institute of General Medical Sciences and K01MH094406 (principal investigator J.J.C.) from the National Institute of Mental Health.

AUTHOR CONTRIBUTIONS

F.P.-M.d.V., J.J.C., L.M., F.Z., W.S., V.Z. and P.F.S. designed the study, and J.J.C. managed the project. J.J.C. and F.P.-M.d.V. drafted the manuscript, and all authors edited it. D.R.M., G.D.S., T.A.B., R.J.B., M.E.C., S.D.H., N.N.R., J.S.S., R.J.N., C.R.Q. and Y.X. bred the mice and collected tissues. J.D.C., C.J.B., Z.Y. and T.J.G. prepared samples for expression profiling. W.S., F.Z., V.Z., Y.K. and W.W. developed statistical models and conducted analyses. W.V., A.B.L., D.W.T., L.M.T., K.K., J.X., J.P.D., A.P.M. and D.L.A. contributed to data analysis and interpretation. S.H., I.K.P., J.R.W., C.E.W., C.-P.F., Z.Z., J.H., Z.G. and L.M. contributed to pseudogenome construction and RNA-seq read alignment.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

2. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
3. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
4. Keane, T.M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
5. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
6. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
7. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
8. Nicolae, D.L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
9. Wang, X. *et al.* Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS ONE* **3**, e3839 (2008).
10. Gregg, C. *et al.* High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science* **329**, 643–648 (2010).
11. Gregg, C., Zhang, J., Butler, J.E., Haig, D. & Dulac, C. Sex-specific parent-of-origin allelic expression in the mouse brain. *Science* **329**, 682–685 (2010).
12. DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genet.* **8**, e1002600 (2012).
13. Wang, X., Soloway, P.D. & Clark, A.G. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. *Genetics* **189**, 109–122 (2011).
14. Okae, H. *et al.* Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. *Hum. Mol. Genet.* **21**, 548–558 (2012).
15. Goncalves, A. *et al.* Extensive compensatory *cis-trans* regulation in the evolution of mouse gene expression. *Genome Res.* **22**, 2376–2384 (2012).
16. Babak, T. *et al.* Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* **18**, 1735–1741 (2008).
17. Hayden, E.C. RNA studies under fire. *Nature* **484**, 428 (2012).
18. Barlow, D.P. Gametic imprinting in mammals. *Science* **270**, 1610–1613 (1995).
19. Skarnes, W.C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).
20. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190**, 389–401 (2012).
21. Churchill, G.A., Gatti, D.M., Munger, S.C. & Svenson, K.L. The Diversity Outbred mouse population. *Mamm. Genome* **23**, 713–718 (2012).
22. Huang, S., Holt, J., Kao, C.Y., McMillan, L. & Wang, W. A novel multi-alignment pipeline for high-throughput sequencing data. *Database (Oxford)* **2014**, bau057 (2014).
23. Zhang, Z. *et al.* GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference due to RNAseq reads misalignment. *Bioinformatics* **29**, 291–299 (2013).
24. Zou, F. *et al.* A novel statistical approach for jointly analyzing RNA-Seq data from F_1 reciprocal crosses and inbred lines. *Genetics* **197**, 389–399 (2014).
25. Wright, F.A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
26. Kim, Y. *et al.* A meta-analysis of gene expression quantitative trait loci in brain. *Transl. Psychiatry* **4**, e459 (2014).
27. Yang, H. *et al.* Subspecific origin and haplotype diversity in the laboratory mouse. *Nat. Genet.* **43**, 648–655 (2011).
28. Xie, W. *et al.* Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**, 816–831 (2012).
29. Ohno, S., Kaplan, W.D. & Kinosita, R. Formation of the sex chromatin by a single X-chromosome in liver cells of *Rattus norvegicus*. *Exp. Cell Res.* **18**, 415–418 (1959).
30. Lyon, M.F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
31. Ohno, S. *Sex Chromosomes and Sex Linked Genes* (Springer Verlag, 1967).
32. Cattanach, B.M. Controlling elements in the mouse X-chromosome. 3. Influence upon both parts of an X divided by rearrangement. *Genet. Res.* **16**, 293–301 (1970).
33. Calaway, J.D. *et al.* Genetic architecture of skewed X inactivation in the laboratory mouse. *PLoS Genet.* **9**, e1003853 (2013).
34. Aylor, D.L. *et al.* Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res.* **21**, 1213–1222 (2011).
35. Cui, X., Affourtit, J., Shockley, K.R., Woo, Y. & Churchill, G.A. Inheritance patterns of transcript levels in F_1 hybrid mice. *Genetics* **174**, 627–637 (2006).
36. Price, A.L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
37. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
38. Kong, A. *et al.* Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
39. Drost, J.B. & Lee, W.R. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila*, mouse, and human. *Environ. Mol. Mutagen.* **25** (suppl. 26), 48–64 (1995).
40. Lin, H. *et al.* Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. *PLoS Biol.* **5**, e326 (2007).
41. Johnston, C.M. *et al.* Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. *PLoS Genet.* **4**, e9 (2008).
42. Yang, F., Babak, T., Shendure, J. & Disteche, C.M. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res.* **20**, 614–622 (2010).
43. Li, N. & Carrel, L. Escape from X chromosome inactivation is an intrinsic property of the *Jarid1c* locus. *Proc. Natl. Acad. Sci. USA* **105**, 17055–17060 (2008).
44. Lopes, A.M. *et al.* Transcriptional changes in response to X chromosome dosage in the mouse: implications for X inactivation and the molecular basis of Turner Syndrome. *BMC Genomics* **11**, 82 (2010).
45. Carrel, L. & Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005).
46. Berletch, J.B., Yang, F. & Disteche, C.M. Escape from X inactivation in mice and humans. *Genome Biol.* **11**, 213 (2010).
47. Nguyen, D.K. & Disteche, C.M. Dosage compensation of the active X chromosome in mammals. *Nat. Genet.* **38**, 47–53 (2006).
48. Gupta, V. *et al.* Global analysis of X-chromosome dosage compensation. *J. Biol.* **5**, 3 (2006).
49. Xiong, Y. *et al.* RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet.* **42**, 1043–1047 (2010).
50. Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat. Genet.* **43**, 1179–1185 (2011).
51. Kharchenko, P.V., Xi, R. & Park, P.J. Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat. Genet.* **43**, 1167–1169 author reply 1171–1172 (2011).
52. Lin, H. *et al.* Relative overexpression of X-linked genes in mouse embryonic stem cells is consistent with Ohno's hypothesis. *Nat. Genet.* **43**, 1169–1170 author reply 1171–1172 (2011).
53. Yildirim, E., Sadreyev, R.I., Pinter, S.F. & Lee, J.T. X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nat. Struct. Mol. Biol.* **19**, 56–61 (2012).
54. He, X. *et al.* He *et al.* reply. *Nat. Genet.* **43**, 1171–1172 (2011).
55. Lin, F., Xing, K., Zhang, J. & He, X. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. *Proc. Natl. Acad. Sci. USA* **109**, 11752–11757 (2012).
56. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
57. Disteche, C.M. Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* **46**, 537–560 (2012).
58. Jue, N.K. *et al.* Determination of dosage compensation of the mammalian X chromosome by RNA-seq is dependent on analytical approach. *BMC Genomics* **14**, 150 (2013).

ONLINE METHODS

Ethical statement. All mouse work was conducted in compliance with the *Guide for the Care and Use of Laboratory Animals* (Institute of Laboratory Animal Resources, National Research Council, 1996) and approved by the Institutional Animal Care and Use Committee of the University of North Carolina.

Mice. The mice used in this study were inbred and reciprocal F_1 hybrids of the wild-derived strains CAST/EiJ, PWK/PhJ and WSB/EiJ. All mice were bred at the University of North Carolina from mice that were fewer than six generations removed from founders acquired from the Jackson Laboratory. Mice were maintained on a 14-h light, 10-h dark schedule with lights turned on at 6 a.m. The housing room was maintained at 20–24 °C with 40–50% relative humidity. Mice were housed in standard 20 cm × 30 cm ventilated polysulfone cages with laboratory-grade Bed-O-Cob bedding. Water and Purina Prolab RMH3000 were available *ad libitum*. A small section of PVC pipe and nestlet material were present in each cage for enrichment.

Tissue collection. Mice were killed at 23 ± 1 d of age by cervical dislocation without anesthesia (to avoid confounding effects on gene expression). All mice were euthanized between 10 and 12 a.m., immediately after removal from their home cage. Whole brain, liver (left lobe), kidneys (both) and lungs (all lobes) were rapidly dissected, snap frozen in liquid nitrogen and pulverized using a BioPulverizer unit (BioSpec Products).

RNA extraction. Total RNA was extracted from ~25 mg of powdered tissue using automated instrumentation (Maxwell 16 Tissue LEV Total RNA Purification Kit, Promega). RNA concentration was measured by fluorometry (Qubit 2.0 Fluorometer, Life Technologies), and RNA quality was verified using a microfluidics platform (Bioanalyzer, Agilent Technologies).

RNA-seq: sample preparation. The 96 samples were randomized to batches of 48 for library preparation using the Illumina TruSeq RNA Sample Preparation Kit v2 with 12 unique indexed adaptors (AD001–AD012). One microgram of total RNA per sample was used as input, and each sample was assigned at least two different barcodes. Libraries were quantified using fluorometry, and 12 randomly selected samples were pooled at equimolar concentrations before sequencing, yielding a total of eight multiplexed pools. The Illumina HiSeq 2000 was used to generate 100-bp paired-end reads. To account for lane and machine effects in cluster density and sequence quality, each sample was divided into four portions, and each portion was randomly assigned to one lane of one machine. The 384 portions (4 × 96 samples) can be partitioned into 18 groups (3 × 3 × 2) for each combination of paternal strain, maternal strain and sex. χ^2 tests confirmed no significant associations between these group indicators and assignments of barcodes or to sequencing lanes.

RNA-seq: alignment. We developed a customized RNA-seq alignment pipeline for mouse subspecies containing considerable genetic diversity[22–24]. This approach has the advantage of incorporating all known strain-specific genetic variants into the alignment reference sequence to improve alignment quality and to minimize bias caused by differences in genetic distance between the parental genomes and the reference sequence. First, reads from each F_1 hybrid (six of the nine cells in the diallel) were aligned to the appropriate 'pseudogenomes' (ref. 22), representing each of the parental genomes using TopHat (v1.4; default parameters including segment length of 25 bp, two mismatches allowed per segment, two mismatches allowed per 100-bp read and a maximum indel of 3 bases). Pseudogenomes were approximations constructed by incorporating all known SNPs and indels into the C57BL/6 genome (mm9). We included all variants reported by a large-scale sequencing effort[4] that included CAST/EiJ, PWK/PhJ and WSB/EiJ (June 2011 release). Second, we mapped coordinates from the pseudogenome-aligned reads to mm9 coordinates by updating the alignment positions and rewriting the CIGAR strings of each aligned read (which was necessary as indels alter the pseudogenome coordinates relative to mm9). Third, we annotated each aligned read to indicate the numbers of maternal and paternal alleles (SNPs and indels) observed in a given read and its paired-end mate. Considering paired-end mates allowed the use of more paired-end reads determining ASE. Finally, alignments to maternal and paternal pseudogenomes were merged by computing the proper union of the separate alignments (i.e., the two alignments were combined such that a read aligning to the same position in both alignments was counted once). This final step was applied separately to all the lanes of a sample, and the resulting alignment files were combined into a single alignment file. For inbred mice, only a single pseudogenome alignment was necessary, followed by the same remapping and annotation stages.

After alignment, we performed a series of quality control checks capitalizing on expectations for the proportions of reads that should align to each parental strain for the sex chromosomes, autosomes and mitochondrial genome. Ninety samples passed quality control.

RNA-seq: read assignment. Three counts were obtained for each gene assessed with RNA-seq: the total number of paired-end reads (for both inbred and F_1 mice; total read count, or TReC) and the numbers of paternal and maternal allele-specific paired-end reads (only for F_1 hybrids; allele-specific read count, or ASReC). A paired-end read was allele specific if either end overlapped at least one SNP or indel that was heterozygous between the paternal and maternal strains. If a paired-end read overlapped more than one heterozygous SNP or indel, it was assigned to a parent only if it was fully consistent (all alleles reported were from one parent and none were from the other). We then counted the number of reads mapped to a gene as the number of paired-end reads that overlapped the exonic regions of a gene using the R function isoform/countReads. Exon position information was assigned on the basis of transcriptome annotations from Ensembl (Release 66, based on mm9; accessed 14 February 2012). There was no need to correct for gene length because all analyses were gene specific and gene length was thus constant in comparisons of the expression of that gene across samples. We included the total number of reads for each sample as a covariate.

RNA-seq: statistical analysis. Statistical analysis is described in detail in Zou *et al.*[24] as well as in the **Supplementary Note**.

Microarray: processing and quality control. Brain, liver, kidney and lung RNA from the same mice used for RNA-seq was hybridized to Affymetrix Mouse Gene 1.1 ST 96-Array Plate arrays using a GeneTitan instrument from Affymetrix according to the manufacturer's protocols. We used the robust multiarray average method (RMA) implemented in the Affymetrix gene expression console with default settings (median polish and sketch-quantile normalization) to estimate the normalized expression levels of transcripts. During normalization, we masked 78,632 probes (~10% of all probes) containing any known SNPs in the 3 mouse inbred strains[4]. We used 28,310 probe sets after excluding control probe sets and those without mRNA annotation. To evaluate the overall performance of the arrays, we applied hierarchical clustering using the R function hclust with the average link function and principal-component analysis (PCA). For inbred strains and reciprocal F_1 crosses between the inbred strains, we fitted linear fixed-effect models for each transcript to test for strain, parent-of-origin, dominance and sex effects (full details are provided below).

Microarray: statistical analysis. For inbred strains and reciprocal F_1 crosses between the inbred strains, we fitted linear fixed-effect models for each transcript to test for strain, parent-of-origin, dominance and sex effects as follows:

$$\begin{aligned} y = {} & \beta_0 + \beta_1\ \text{strain} + \beta_2\ \text{parent of origin} + \beta_3\ \text{dominance} + \beta_4\ \text{sex} \\ & + \beta_5\ \text{strain} \times \text{sex} + \beta_6\ \text{parent of origin} \times \text{sex} + \beta_7\ \text{dominance} \times \text{sex} \\ & + \beta_8\ \text{plate} + \beta_9\ \text{dissection} + \varepsilon \end{aligned}$$

where "strain" is a vector for comparisons of two inbred strains, "parent of origin" is a vector for comparisons of reciprocal F_1 crosses, "dominance" indicates reciprocal F_1 crosses, "sex" indicates female, "plate" is a categorical variable indicating multiple 96-well plates and "dissection" is a categorical variable

indicating different dissection dates. We tested the strain, parent-of-origin, dominance and sex effects as follows:

Strain effect: H_0: $\beta_1 = \beta_5 = 0$ vs. H_1: $\beta_1 \neq 0$ or $\beta_5 \neq 0$

Parent-of-origin effect: H_0: $\beta_2 = \beta_6 = 0$ vs. H_1: $\beta_2 \neq 0$ or $\beta_6 \neq 0$

Dominance effect: H_0: $\beta_3 = \beta_7 = \mathbf{0}$ vs. H_1: $\beta_3 \neq 0$ or $\beta_7 \neq 0$

Sex effect: H_0: $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ vs. H_1: $\beta_4 \neq 0$ or $\beta_5 \neq 0$ $\beta_6 \neq 0$ or $\beta_7 \neq 0$

For multiple-testing correction, we used false discovery rate (FDR) and declared tests to be significant if the q value was <0.05.

Paternal expression bias in RNA-seq data. To quantify the paternal expression bias shown in **Figure 6a**, we permuted a random subset of 1,000 genes (minus known imprinted genes) 2,000 times. We used a random subset of genes to avoid P-value inflation due to possible correlation between genes (this test is therefore conservative yet still highly significant). For each random subset of 1,000 genes, we tested whether the expected paternal expression proportion was different from 50% using a Wilcoxon rank-sum test (we used the median result from 2,000 simulations). These tests were performed separately for each combination of cross and sex (and were significant in each case) and then collapsed into one P value using Fisher's combined probability test.

This parent-of-origin effect on allelic imbalance was observed on every autosome, and there was evidence of clustering. To quantify the magnitude of the clustering shown in **Figure 6b**, we performed the following procedure. For each cross and each sex, we checked whether neighboring genes had the same direction of parent-of-origin effect. We recorded the proportion of such genes within each chromosome after pooling results from three crosses and both sexes. Then we compared these chromosome-wise proportions with what would be expected under the null: $p^2 + (1 - p)^2$, where p is the proportion of paternally overexpressed genes for the corresponding chromosome. We found that, for 15 of 19 autosomes, the observed proportion was higher than expected.

To further explore this clustering, we calculated the distance from the TSS to the nearest CpG island for all 467 genes that were consistently paternally overexpressed and all 116 genes that were consistently maternally overexpressed. We compared these distances to those for the remainder of expressed genes (with inconsistent parental expression) to generate respective distributions. Paternally overexpressed genes tended to be closer to CpG islands than inconsistent genes, and maternally overexpressed genes tended to be further away from CpG islands (**Fig. 6c**). To formally test the significance of this difference, we randomly sampled 467 and 116 genes from the whole gene list and calculated the mean squared deviation of the curves. We repeated this procedure 100,000 times and calculated the P value as the proportion of times where the mean squared deviation from randomly sampled genes was larger than the one from unperturbed data. The resulting P values were $<1 \times 10^{-5}$ for paternally overexpressed genes (out of 100,000 permutations, none was as extreme as the empirical result) and 1×10^{-5} for maternally overexpressed genes.

Relationship between paternal expression bias and DNA methylation. We tested whether genes with consistent overexpression from the paternal allele were closer to CpG islands with parent-of-origin bias in methylation (**Supplementary Fig. 10**). To accomplish this, we used a data set (GSE33722, Gene Expression Omnibus) from a recent publication by Xie *et al.*[28]. This data set consists of whole-genome parent-of-origin brain DNA methylation data from reciprocal hybrids of 129X1/SvJ and Cast/EiJ mice. Because this data set included just one mouse per reciprocal cross, we first integrated CpG methylation counts over each CpG island and applied a simple filter criterion: if both mice had a maternal methylation proportion higher than the paternal proportion, we declared this CpG island to be preferentially maternally methylated, for the purposes of this analysis. Likewise, if both mice had a paternal methylation proportion higher than the maternal proportion, we declared the CpG island to be preferentially paternally methylated. The remaining CpG islands with no preferential methylation were used as a reference group.

Next, we calculated the distance from each gene's TSS to the closest CpG island for each parentally biased methylation group and examined the distribution of these distances with respect to parental overexpressed. In other words, we examined distributions for all combinations of methylation group (maternal, paternal and other) and overexpressed group (paternal and maternal), six combinations in total. To avoid bias due to differential CpG island count per group, we calculated distance to a down-sampled subset equivalent to the smallest group, and to make the result more robust we used 10,000 permutations of the median distance between the TSS and the closest CpG island. A comparison of consistently paternally overexpressed genes and inconsistently expressed genes, using the following log ratio: $\log_{10}$ (paternally expressed: TSS to nearest CpG island (bp)/inconsistently expressed: TSS to nearest CpG island (bp)) is shown in **Supplementary Figure 10**. In short, this plot examines whether consistent paternally overexpressed genes tended to be closer than inconsistent genes to each class of CpG island. We found that paternally overexpressed genes had the greatest enrichment for maternally methylated CpG islands (permutation $P = 0$), followed by paternally methylated CpG islands ($P = 0.0034$). This greater enrichment for maternal over paternal methylated CpG islands was itself also significant ($P = 0.0015$).

doi:10.1038/ng.3222

Corrigendum: Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance

James J Crowley, Vasyl Zhabotynsky, Wei Sun, Shunping Huang, Isa Kemal Pakatci, Yunjung Kim, Jeremy R Wang, Andrew P Morgan, John D Calaway, David L Aylor, Zaining Yun, Timothy A Bell, Ryan J Buus, Mark E Calaway, John P Didion, Terry J Gooch, Stephanie D Hansen Nashiya N Robinson, Ginger D Shaw, Jason S Spence, Corey R Quackenbush, Cordelia J Barrick, Randal J Nonneman, Kyungsu Kim, James Xenakis, Yuying Xie, William Valdar, Alan B Lenarcic, Wei Wang, Catherine E Welsh, Chen-Ping Fu, Zhaojun Zhang, James Holt, Zhishan Guo, David W Threadgill, Lisa M Tarantino, Darla R Miller, Fei Zou, Leonard McMillan, Patrick F Sullivan & Fernando Pardo-Manuel de Villena

***Nat. Genet.* 47, 353–360 (2015); published online 2 March 2015; corrected after print 16 April 2015**

In the version of this article initially published, an accession number was not provided for RNA-seq data sets. The RNA-seq data sets that passed quality control are available at the Sequence Read Archive (SRA) under accession SRP056236. The error has been corrected in the HTML and PDF versions of the article.