UCLA
UCLA Previously Published Works

Title

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

Permalink

https://escholarship.org/uc/item/9vw3r2qw

Journal

Nature Genetics, 47(4)

ISSN

1061-4036

Authors

Crowley, James J
Zhabotynsky, Vasyl
Sun, Wei
et al.

Publication Date

2015-04-01

DOI

10.1038/ng.3222

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2015 October 01.

Published in final edited form as:

Nat Genet. 2015 April ; 47(4): 353–360. doi:10.1038/ng.3222.

Analyses of Allele-Specific Gene Expression in Highly Divergent Mouse Crosses Identifies Pervasive Allelic Imbalance

James J Crowley[1,11], **Vasyl Zhabotynsky**[1,11], **Wei Sun**[1,2,11], **Shunping Huang**[3], **Isa Kemal Pakatci**[3], **Yunjung Kim**[1], **Jeremy R Wang**[3], **Andrew P Morgan**[1], **John D Calaway**[1], **David L Aylor**[1], **Zaining Yun**[1], **Timothy A Bell**[1], **Ryan J Buus**[1], **Mark E Calaway**[1], **John P Didion**[1], **Terry J Gooch**[1], **Stephanie D Hansen**[1], **Nashiya N Robinson**[1], **Ginger D Shaw**[1], **Jason S Spence**[1,9], **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**[3,10], **Catherine E Welsh**[3], **Chen-Ping Fu**[3], **Zhaojun Zhang**[3], **James Holt**[3], **Zhishan Guo**[3], **David W Threadgill**[5], **Lisa M Tarantino**[6], **Darla R Miller**[1], **Fei Zou**[2,12], **Leonard McMillan**[3,12], **Patrick F Sullivan**[1,6,7,8,12], and **Fernando Pardo-Manuel de Villena**[1,4,7,12]

[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

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Correspond with: Fernando Pardo-Manuel de Villena, PhD, Department of Genetics, University of North Carolina, Chapel Hill, NC 27599-7264. Tel: 919-843-5403. Fax: 919-843-4682. fernando@med.unc.edu.

[9]Present address: Department of Animal Science, University of Tennessee, Knoxville, Tennessee, USA.

[10]Present address: Department of Computer Science, University of California, Los Angeles, California, USA.

[11]These authors contributed equally to this work.

[12]These authors jointly directed this work.

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 total and allele-specific reads counts, and factor in X inactivation skewing, can be found at www.bioconductor.org ("rxSeq"). To detect and correct spurious RNAseq read misalignment (i.e. pseudogenes) access GeneScissors (http://csbio.unc.edu/genescissors/). Knockout mouse phenotypes were acquired from: www.informatics.jax.org/phenotypes.shtml. Orthologous genes between human and mice were identified from Ensembl (www.ensembl.org/info/genome/compara/homology_method.html) using the category "ortholog_one2one". Genes with prior evidence of imprinting were identified by creating a union of the following databases: www.geneimprint.com, http://igc.otago.ac.nz, www.mousebook.org/catalog.php?catalog=imprinting.

Accession Codes

Expression data can be acquired from Gene Expression Omnibus (GEO) accession ID GSE44555.

Author Contributions

F.P-M.dV., 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.dV. 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.F., Z.Z., J.H., Z.G., L.M. contributed to pseudogenome construction and RNAseq read alignment.

Competing Financial Interests

The authors declare no competing financial interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[4]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[5]Department of Molecular and Cellular Medicine, Texas A&M Health Science Center, College Station, Texas, USA

[6]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[7]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[8]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Abstract

Complex human traits are influenced by variation in regulatory DNA through mechanisms that are not fully understood. Since regulatory elements are conserved between humans and mice, a thorough annotation of *cis* regulatory variants in mice could aid in this process. 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. These effects 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 novel, global allelic imbalance in favor of 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.

Keywords

mouse; eQTL; allelic imbalance; imprinting; dosage compensation

Introduction

The genetic basis of most phenotypic variation can be assigned to variation in protein, RNA, or regulatory sequences. The significance of regulatory sequence has become increasingly apparent in recent studies comparing divergent taxa and populations[1–4] and by 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. 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 in order to detect and quantify statistically significant allelic imbalance in as many genes as possible.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

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). These strains were chosen to maximize the level of genetic diversity (e.g., 27.7 million SNPs and 4.6 million indels that vary in these strains[4]), the number of genes with expressed SNPs and/or indels (31,259 out of 36,817 Ensembl v37 genes), and the number of such variants per gene (mean 19.9, standard deviation 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. RNAseq was used to measure allele-specific expression in brain and microarrays were used to assess gene expression in brain, liver, kidney and lung. Inclusion of the array data allowed a detailed comparison of two major expression platforms, 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 like the International Knockout Mouse Consortium,[19] Collaborative Cross[20] and Diversity Outbred.[21] We also increased the depth of sequencing, the number of replicates and included both sexes in order to improve power to detect ASE. We developed a novel 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 an F_1 animal requires the presence of a genetic or epigenetic regulatory variant acting in *cis*, because *trans* acting factors have 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 ASE derived from each parental allele in an F_1 animal (Supplementary Fig. 2). We observe *cis* regulatory effects in >85% of all testable genes. We also find that the number of imprinted

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genes is not substantially different from historical estimates, but we report a new genome-wide parent-of-origin allelic imbalance favoring the paternal allele.

Results

Major drivers of differential gene expression in mice

Brain, liver, kidney and lung RNA from the same mice used for RNAseq were hybridized 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, 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 revealed 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 RNAseq total read counts and microarray intensity values were highly correlated (median $r = 0.86$, range 0.84–0.87).

Within each tissue, the overwhelming driver of differential gene expression was strain; this greatly exceeded the effects of parent-of-origin and sex (Fig. 2). For RNAseq, the first two principal components (PCs) accounted for ~30% of the total variation in autosomal total read count (TReC). The remaining top 10 PCs were also strongly determined by strain and, to a far lesser extent, parent-of-origin and sex, with no notable effects of the barcodes used for multiplexing (Supplementary Table 3).

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

Cis-regulatory variation is pervasive in diverse mice

Cis regulatory effects were found in 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* shows allelic imbalance in 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_1s 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, due to dominance and other effects. For example, *Fos* shows allelic imbalance in all F_1s in a manner consistent with total read counts in the parental inbreds, but the total level of gene expression in the F_1s is best explained as an effect of dominance or over-dominance (Supplementary Fig. 5). Copy number variation can also lead to inconsistency

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

between TReC and ASReC and result in underestimation of genes with *cis* effects (see Discussion).

Of the 11,287 autosomal genes with *cis* regulatory effects, 4,113 (36%) were detected between all three pairs of strains, 5,065 (45%) between two pairs, and 2,109 (19%) between one pair (Fig. 3a). Critically, all three subspecies contribute 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 up- to down-regulated genes in any cross (Fig. 3b). A similar pattern was seen with the microarray data across four tissues (Supplementary Fig. 6).

Phenotypic consequences and human relevance

To test the potential significance of *cis* regulatory variation, we compared our results to a comprehensive set of knockout mice phenotypes for 6,039 different genes and 29 phenotype dimensions (see URLs). Brain expressed genes with *cis* regulatory effects were significantly more likely to cause a behavioral or neurological phenotype in knockout mice ($P = 0.012$), relative to brain expressed genes with no *cis* effect. Furthermore, no such enrichment was found for the 1,348 genes that result in no overt 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 human eQTL studies. These comparisons were restricted to only those genes that have a one-to-one ortholog between mouse and human ($n = 15,312$ genes, see URLs). Brain expressed genes with a *cis* regulatory effect in mouse were much more likely to possess a human peripheral blood eQTL ($P = 7.8 \times 10^{-10}$).[25] Published human brain eQTL studies have much smaller samples sizes, nonetheless, after comparing our results to a meta-analysis[26] of five available datasets (total $n = 439$), we observe a consistent enrichment ($P = 0.04$).

Proportion of SNPs with *cis* regulatory effects

In contrast to previous F_1 RNAseq studies, we included three genomes in our experimental design in order to allow multiple pairwise comparisons. In our experiment, for >90% of the genome, pairwise comparisons are possible between different subspecies (*domesticus*, *musculus* or *castaneus*), while for the remainder of the genome, just one subspecies is represented (*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 show 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 (DGE) (Supplementary Fig. 7) and this replicated in all four tissues.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Each *cis* eQTL identified in this study is caused by at least one regulatory variant. Therefore, we can 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 is 0.10% (±0.02%) so that approximately 1 in 1,000 SNPs create a *cis* regulatory effect. This estimate was stable across all crosses examined, and all regions independently of their phylogenetic origin. This estimate also generalized to genes of varying size and level of expression.

Classical imprinting is incomplete and under genetic control

We identified 95 genes with significant imprinting (Fig. 5a, full gene list in Supplementary Dataset). Imprinted genes were found on 16 chromosomes, with 62 of these 95 genes residing in well-known imprinting clusters (Supplementary Fig. 9). There were 52 novel imprinted genes and 43 genes with prior evidence of imprinting (see URLs). Of 128 genes with prior evidence of imprinting from the literature, 73 genes could be evaluated (expressed and containing exonic variation) and 42 genes (58%) were imprinted. The remaining 31 unidentified genes were sufficiently expressed (median TReC 809, median ASReC 143) but did not meet 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 RNAseq data allows quantification of the strength of imprinting in each gene. For most genes, imprinting is incomplete. In maternally expressed genes, maternal reads represent an average 67% of ASReC (range: 51.5% – 97.9%). In paternally expressed genes, paternal reads represent an average 75.6% of ASReC (range: 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 show a strain effect modifying the strength of imprinting (strain by parent-of-origin effect). We divided these 47 genes into two classes: those in which we can explain the differential gene expression based on a single strain effect ($n = 11$) and those where we cannot, suggesting 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 observation is consistent with the observation that paternal expression predominates in brain, while maternal expression predominates in placenta.[9] To test whether this asymmetry in parent-of-origin effects extends beyond imprinted genes, we estimated the parent-of-origin effect in each cross and each sex separately. We found that 54–60% of genes show higher expression from the paternal allele, significantly different from the expectation of 50% ($P = 5.9 \times 10^{-24}$, Fig. 6a, Supplementary Table 5). We also observed that genes with higher expression from one parental allele tend to cluster (Fig. 6b). Among the 19 autosomes, 15 have a higher proportion of genes whose neighbor has the same parental skew than expected by chance ($P = 9.6 \times 10^{-3}$, binomial test).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We can calculate a rough estimate of the number of genes with paternal overexpression, simply by taking the difference between the number of genes with higher paternal minus higher maternal expression. For example, for female CAST/EiJ × PWK/PhJ reciprocal hybrids, there are 1,652 more genes with allelic imbalance in favor of the paternal allele (6,790 paternal minus 5,138 maternal overexpressed genes). As shown in Supplementary Table 5, the excess of genes with paternal overexpression ranges between 938 and 2,500 (across reciprocal crosses stratified by sex). However, this likely represents an underestimate because, while we have high power to identify classical imprinting (Fig. 5a), 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 are 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 (TSS) of genes with consistent overexpression of the paternal allele in all three crosses (N = 467 with and 3,338 without stratification by sex) are closer to CpG islands ($P < 1\times10^{-5}$) relative to the remaining genes (Fig. 6c, 6d). This effect is not observed among maternally consistent genes (N = 116 and 1,631, $P = 0.60$). Note that for the more restrictive group (consistently expressed in both sexes within each cross), there is a further enrichment for genes with TSS near CpG islands among paternally overexpressed and a significant depletion for genes with TSS near CpG islands among maternally consistent genes (Fig. 6c, $P = 1 \times 10^{-5}$).

For genes consistently overexpressing the paternal allele, we observe that the size of the strain effect is 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, proximity of a CpG island to the TSS is associated with smaller additive strain effect sizes, and genes with TSS that overlap CpG islands are also clustered in the genome. We conclude that, in addition to statistically significant allelic imbalance observed at the gene level (imprinting), there is an association between proximity of a CpG island to TSS and a pervasive allelic imbalance favoring 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 dataset from reciprocal hybrids of 129X1/SvJ and CAST/EiJ mice.[28] Genes with consistent overexpression from the paternal allele are closer to CpG islands that are 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 paternal versus maternal methylation bias.[28]

Two forms of dosage compensation on chromosome X

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

autosomal genes[31]. In our dataset, the overall level of chromosome X gene expression is 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, chromosome X gene expression is equivalent to the autosomes in all four tissues examined (Supplementary Figs. 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 their Y chromosome homologs. We also observed an effect of X chromosome controlling element (*Xce*[32]) genotype and a parent-of-origin effect in X chromosome inactivation skewing in females (Supplementary Fig. 12).[33]

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

Discussion

We find that more than 80% of mouse genes have expression levels dependent upon 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. Differential gene expression is positively correlated with sequence diversity at multiple evolutionary scales, and the proportion of mutations that create a *cis* regulatory effect remained relatively constant as mouse subspecies evolved. We observe two types of parent-of-origin effects on gene expression. We demonstrate that the number of imprinted genes is not substantially different from historical estimates. We also observe a global allelic imbalance in favor 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 imprint. Furthermore, we conclude that regulation of gene expression on chromosome X is similar to the autosomes and includes two forms of dosage compensation. Finally, we developed improved analytical tools with broad utility for RNA sequencing 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 RNAseq read misalignment and allow analysis of expression on chromosome X without chromosome-wide confounding effects. Finally, a novel likelihood-based method to jointly analyze TReC and ASE from inbred and F_1 mice (Supplementary Fig. 2) increases statistical power to detect genetic effects.

Cis regulatory effects were found in 11,686 genes (85% of testable genes). This number exceeds all prior mouse eQTL studies.[34] We found that the expression of most transcripts show an additive pattern of inheritance, consistent with mouse,[35] human[36] and plant[37] studies. Interestingly, many genes have inconsistent patterns of inheritance between TReC

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and ASE. 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 RNAseq study, we identified one XO female caused by paternal nondisjunction (supported by genotyping) and another mouse with a ~250 kb duplication spanning five genes (Supplementary Fig. 13).

Pinpointing the genetic variants that underlie mouse quantitative trait loci has been challenging because QTL detected in experimental crosses often span hundreds of genes. The data described here can help investigators prioritize candidate genes based on 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. Therefore, animal models for such regulatory variation are needed to provide a more detailed understanding of genotype to 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, since 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 imprinting is restricted to genes that show significant allelic imbalance toward one parent, then our results indicate that the number of genes imprinted in mouse brain is in line with the historical consensus. On the other hand, 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 in every autosome and is present in all three

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

reciprocal crosses. The 467 genes that have consistent overexpression of the paternal allele in all three crosses and both sexes are strongly enriched for CpG islands near their TSS 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 strain effect size between genes that are effected 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 the overall methylation. Lastly, this global allelic imbalance in favor of the paternal allele may partly explain why the majority of the novel imprinted genes described here (37 of 54) show modest overexpression of the paternal allele and also the surprisingly large number of genes found in two previous controversial studies.[10,11]

We verified two forms of dosage compensation on the X chromosome. First, for most of the genes on X, 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 live mice. Furthermore, we confirm that it is rare for genes to escape X inactivation in mouse, with this occurring in just 1.1 % of genes that could be tested all of which having been previously identified.[42–44] This stands in sharp contrast to 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 RNAseq 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 to include genes with low expression.[57,58] Since genes with no or low expression in somatic tissues are more abundant on X than autosomes,[50] inclusion can lower the median X:autosome expression ratios. Our analysis considered all genes on chromosome X and clearly supports Ohno’s hypothesis in *Mus musculus*. 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 mechanisms underlying human complex genetic traits.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Online Methods

Ethical Statement

All animal 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 animals were bred at UNC from mice that were less than six generations removed from founders acquired from the Jackson Laboratory (Bar Harbor, ME). Animals were maintained on a 14-hour light, 10-hour dark schedule with lights on at 0600. 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 sacrificed at 23±1 days of age by cervical dislocation without anesthesia (to avoid confounding effects on gene expression). All mice were euthanized between 10:00 AM and 12:00 PM, 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, Bartlesville, OK).

RNA extraction

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

RNAseq: sample preparation

The 96 samples were randomized to batches of 48 for library preparation using the Illumina (San Diego, CA) TruSeq RNA Sample Preparation Kit v2 with 12 unique indexed adapters (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 quantitated using fluorometry and 12 randomly selected samples were pooled at equimolar concentrations prior to sequencing, yielding a total of 8 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

strain, and sex. Chi-squared tests confirmed no signification associations between these group indicators and assignments of barcodes or to sequencing lanes.

RNAseq: alignment

We developed a customized RNAseq alignment pipeline for mouse sub-species containing considerable genetic diversity.[22–24] This 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 to the reference sequence. First, reads from each F_1 hybrid (six of the nine cells in the diallel) were aligned to the appropriate 'pseudogenomes'[22] representing each of the parental genomes using TopHat (v1.4, default parameters including segment length 25 bp, 2 mismatches allowed per segment, 2 mismatches allowed per 100 bp read, and maximum indel of 3 bases). Pseudogenomes are 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 (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 the 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.

Following 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.

RNAseq: read assignment

Three counts were obtained for each gene assessed with RNAseq: 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/indel, it was assigned to a parent only if it was fully consistent (all alleles reported were from one parent and zero were from the other). We then counted the number of reads mapped to a gene as the number of paired-end reads that overlapped exonic regions of a gene using the R function isoform/countReads. Exon position information was assigned based on transcriptome annotation from Ensembl (Release 66 based on mm9, accessed 2/14/2012). There was no need to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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 total number of reads per sample as a covariate.

RNAseq: statistical analysis

This is described in detail in Zou et al[24] as well as the Supplementary Note.

Microarray: processing and QC

Brain, liver, kidney and lung RNA from the same mice used for RNAseq were 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 robust multiarray average method (RMA) implemented in the Affymetrix gene expression console with default settings (median polish and sketch-quantile normalization) to estimate normalized expression levels of transcripts. During normalization, we masked 78,632 probes (~10% of all probes) containing any known SNPs in these three mouse inbred strains.[4] We used 28,310 probe-sets after excluding control probe-sets and those without mRNA annotation. In order to evaluate overall performance of arrays, we used 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 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 following:

$$y=\beta_0+\beta_1\, strain+\beta_2\, parent-of-orgin+\beta_3\, dominance+\beta_4\, sex+\beta_5\, strain\times sex+\beta_6\, parent-of-origin\times sex+\beta_7\, dominance\times sex+\beta_8\, plate+\beta_9\, dissection+\varepsilon$$

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 catogorical variable indicating multiple 96-well plates and "dissection" is a categorical variable indicating different disecction dates, respectively. We test the strain, parent-origin, dominance, and sex effects as follows:

$$\begin{aligned}
\text{Strain effect:} &\quad H_0{:}\beta_1=\beta_5=0\ vs.\ H_1{:}{:}\beta_1\neq 0\ or\ \beta_5\neq 0\\
\text{Parent- of- origin effect:} &\quad H_0{:}\beta_2=\beta_6=0\ vs.\ H_1{:}{:}\beta_2\neq 0 \text{ or } \beta_6\neq 0\\
\text{Dominance effect:} &\quad H_0{:}\beta_3=\beta_7=0\ vs.\ H_1{:}{:}\beta_3\neq 0\ or\ \beta_7\neq 0\\
\text{Sex effect:} &\quad H_0{:}\beta_4=\beta_5=\beta_6=\beta_7=0\ vs.\ H_1{:}{:}\beta_4\neq 0\ or\ \beta_5\neq 0\ \beta_6\neq 0\ or\ \beta_7\neq 0
\end{aligned}$$

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Paternal expression bias in RNAseq data

To quantify the paternal expression bias shown in Fig. 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 clustering shown in Fig. 6b, we performed the following procedure. For each cross and each sex, we checked whether neighboring genes have the same direction of parent-of-origin effect. We recorded of 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, in 15 out of 19 autosomes, the observed proportion was higher than expected.

To further explore this clustering, we calculated the distance from the transcription start site (TSS) to the nearest CpG island for all 467 genes that were consistently paternally expressed and all 116 genes consistently maternally expressed. We compared these distances to those for the remainder of expressed genes (inconsistent parental expression) to generate respective distributions. As shown in Fig. 6c, paternally expressed genes tend to be closer to CpG islands than inconsistent genes, and maternally expressed genes tend to be further away from CpG islands. To formally test the significance of this difference, we randomly sampled 467 and 116 genes from the whole gene list and calculated 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 $< 10^{-5}$ for paternally expressed genes (i.e. out of 100,000 permutations, none was as extreme as the empirical result) and 10^{-5} for maternally expressed genes.

Relationship between paternal expression bias and DNA methylation

As shown in Supplementary Fig. 10, we tested whether genes with consistent overexpression from the paternal allele were closer to CpG islands with parent-of-origin methylation. To accomplish this, we used a dataset (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33722) from a recent publication by Xie et al.[28] This dataset consists of whole genome parent of origin brain DNA methylation data from reciprocal hybrids of 129X1/SvJ and Cast/EiJ mice. Since this dataset 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 paternal, 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

higher than maternal, we declared it 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 parental methylation group and examined the distribution of these distances with respect to parental expression. In other words, we examined distributions for all combinations of methylation group (maternal, paternal and others) and expression group (paternal and maternal), six combinations in total. In order 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 TSS and the closest CpG island. Supplementary Fig. 10 shows a comparison of consistently paternally expressed genes versus 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]). In short, this plot examines whether consistent paternally expressed genes tend to be closer than inconsistent genes with each class of CpG islands. We find that paternally expressed genes have the greatest enrichment for maternally methylated islands (permutation $P = 0$) followed by paternally methylated islands ($P = 0.0034$). This greater enrichment for maternal over paternal methylated island is itself significant as well ($P = 0.0015$).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Major funding was provided by an NIMH/NHGRI Center of Excellence for Genome Sciences grant (P50MH090338 and P50HG006582, co-PIs Dr. Fernando Pardo-Manuel de Villena and Dr. Patrick F. Sullivan). This work was also supported by R01GM074175 (PI Dr. Fei Zou) and K01MH094406 (PI Dr. James J. Crowley). We thank Drs Piotr Mieczkowski, Alicia Brandt, Ewa Malc, Michael Vernon, Jennifer Brennan and Mauro Calabrese for helpful discussions.

References

1. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
2. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–16. [PubMed: 1090005]
3. Gan X, et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature. 2011; 477:419–23. [PubMed: 21874022]
4. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477:289–94. [PubMed: 21921910]
5. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–5. [PubMed: 22955828]
6. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:9362–7. [PubMed: 19474294]
7. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome research. 2012; 22:1748–59. [PubMed: 22955986]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS genetics. 2010; 6:e1000888. [PubMed: 20369019]
9. Wang X, et al. Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. PloS one. 2008; 3:e3839. [PubMed: 19052635]
10. Gregg C, et al. High-resolution analysis of parent-of-origin allelic expression in the mouse brain. Science. 2010; 329:643–8. [PubMed: 20616232]
11. Gregg C, Zhang J, Butler JE, Haig D, Dulac C. Sex-specific parent-of-origin allelic expression in the mouse brain. Science. 2010; 329:682–5. [PubMed: 20616234]
12. DeVeale B, van der Kooy D, Babak T. Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. PLoS genetics. 2012; 8:e1002600. [PubMed: 22479196]
13. Wang X, Soloway PD, Clark AG. A survey for novel imprinted genes in the mouse placenta by mRNA-seq. Genetics. 2011; 189:109–22. [PubMed: 21705755]
14. Okae H, et al. Re-investigation and RNA sequencing-based identification of genes with placenta-specific imprinted expression. Human molecular genetics. 2012; 21:548–58. [PubMed: 22025075]
15. Goncalves A, et al. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. Genome research. 2012
16. Babak T, et al. Global survey of genomic imprinting by transcriptome sequencing. Current biology : CB. 2008; 18:1735–41. [PubMed: 19026546]
17. Hayden EC. RNA studies under fire. Nature. 2012; 484:428. [PubMed: 22538578]
18. Barlow DP. Gametic imprinting in mammals. Science. 1995; 270:1610–3. [PubMed: 7502071]
19. Skarnes WC, et al. A conditional knockout resource for the genome-wide study of mouse gene function. Nature. 2011; 474:337–42. [PubMed: 21677750]
20. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 2012; 190:389–401. [PubMed: 22345608]
21. Churchill GA, Gatti DM, Munger SC, Svenson KL. The Diversity Outbred mouse population. Mamm Genome. 2012; 23:713–8. [PubMed: 22892839]
22. Huang S, Holt J, Kao CY, McMillan L, Wang W. A novel multi-alignment pipeline for high-throughput sequencing data. Database (Oxford). 2014
23. Zhang, Z., et al. GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference due to RNAseq reads misalignment. Proceedings of the 21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Special Issue of Bioinformatics; 2013;
24. Zou F, et al. A novel statistical approach for jointly analyzing RNA-Seq data from F1 reciprocal crosses and inbred lines. Genetics. 2014; 197:389–99. [PubMed: 24561482]
25. Wright FA, et al. Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014; 46:430–7. [PubMed: 24728292]
26. Kim Y, et al. A meta-analysis of gene expression quantitative trait loci in brain. Transl Psychiatry. 2014; 4:e459. [PubMed: 25290266]
27. Yang H, et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics. 2011; 43:648–55. [PubMed: 21623374]
28. Xie W, et al. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell. 2012; 148:816–31. [PubMed: 22341451]
29. Ohno S, Kaplan WD, Kinosita R. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. Experimental cell research. 1959; 18:415–8. [PubMed: 14428474]
30. Lyon MF. Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature. 1961; 190:372–3. [PubMed: 13764598]
31. Ohno, S. Sex Chromosomes and Sex Linked Genes. Springer Verlag; Berlin: 1967.
32. Cattanach BM. Controlling elements in the mouse X-chromosome. 3. Influence upon both parts of an X divided by rearrangement. Genetical research. 1970; 16:293–301. [PubMed: 5512255]
33. Calaway JD, et al. Genetic architecture of skewed X inactivation in the laboratory mouse. PLoS Genet. 2013; 9:e1003853. [PubMed: 24098153]
34. Aylor DL, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome research. 2011; 21:1213–22. [PubMed: 21406540]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

35. Cui X, Affourtit J, Shockley KR, Woo Y, Churchill GA. Inheritance patterns of transcript levels in F1 hybrid mice. Genetics. 2006; 174:627–37. [PubMed: 16888332]
36. Price AL, et al. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS genetics. 2011; 7:e1001317. [PubMed: 21383966]
37. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422:297–302. [PubMed: 12646919]
38. Kong A, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012; 488:471–5. [PubMed: 22914163]
39. Drost JB, Lee WR. Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among drosophila, mouse, and human. Environmental and molecular mutagenesis. 1995; 25 (Suppl 26):48–64. [PubMed: 7789362]
40. Lin H, et al. Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. PLoS biology. 2007; 5:e326. [PubMed: 18076287]
41. Johnston CM, et al. Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. PLoS genetics. 2008; 4:e9. [PubMed: 18208332]
42. Yang F, Babak T, Shendure J, Disteche CM. Global survey of escape from X inactivation by RNA-sequencing in mouse. Genome research. 2010; 20:614–22. [PubMed: 20363980]
43. Li N, Carrel L. Escape from X chromosome inactivation is an intrinsic property of the Jarid1c locus. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:17055–60. [PubMed: 18971342]
44. Lopes AM, 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. 2010; 11:82. [PubMed: 20122165]
45. Carrel L, Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature. 2005; 434:400–4. [PubMed: 15772666]
46. Berletch JB, Yang F, Disteche CM. Escape from X inactivation in mice and humans. Genome biology. 2010; 11:213. [PubMed: 20573260]
47. Nguyen DK, Disteche CM. Dosage compensation of the active X chromosome in mammals. Nature genetics. 2006; 38:47–53. [PubMed: 16341221]
48. Gupta V, et al. Global analysis of X-chromosome dosage compensation. Journal of biology. 2006; 5:3. [PubMed: 16507155]
49. Xiong Y, et al. RNA sequencing shows no dosage compensation of the active X-chromosome. Nature genetics. 2010; 42:1043–7. [PubMed: 21102464]
50. Deng X, et al. Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. Nature genetics. 2011; 43:1179–85. [PubMed: 22019781]
51. Kharchenko PV, Xi R, Park PJ. Evidence for dosage compensation between the X chromosome and autosomes in mammals. Nature genetics. 2011; 43:1167–9. author reply 1171–2. [PubMed: 22120048]
52. Lin H, et al. Relative overexpression of X-linked genes in mouse embryonic stem cells is consistent with Ohno's hypothesis. Nature genetics. 2011; 43:1169–70. author reply 1171–2. [PubMed: 22120049]
53. Yildirim E, Sadreyev RI, Pinter SF, Lee JT. X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. Nature structural & molecular biology. 2012; 19:56–61.
54. He X, et al. He et al. Reply. Nature Genetics. 2011; 43:1171–1172.
55. Lin F, Xing K, Zhang J, He X. Expression reduction in mammalian X chromosome evolution refutes Ohno's hypothesis of dosage compensation. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:11752–7. [PubMed: 22753487]
56. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–8. [PubMed: 22012392]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

57. Disteche CM. Dosage compensation of the sex chromosomes. Annual review of genetics. 2012; 46:537–60.

58. Jue NK, et al. Determination of dosage compensation of the mammalian X chromosome by RNA-seq is dependent on analytical approach. BMC genomics. 2013; 14:150. [PubMed: 23497106]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

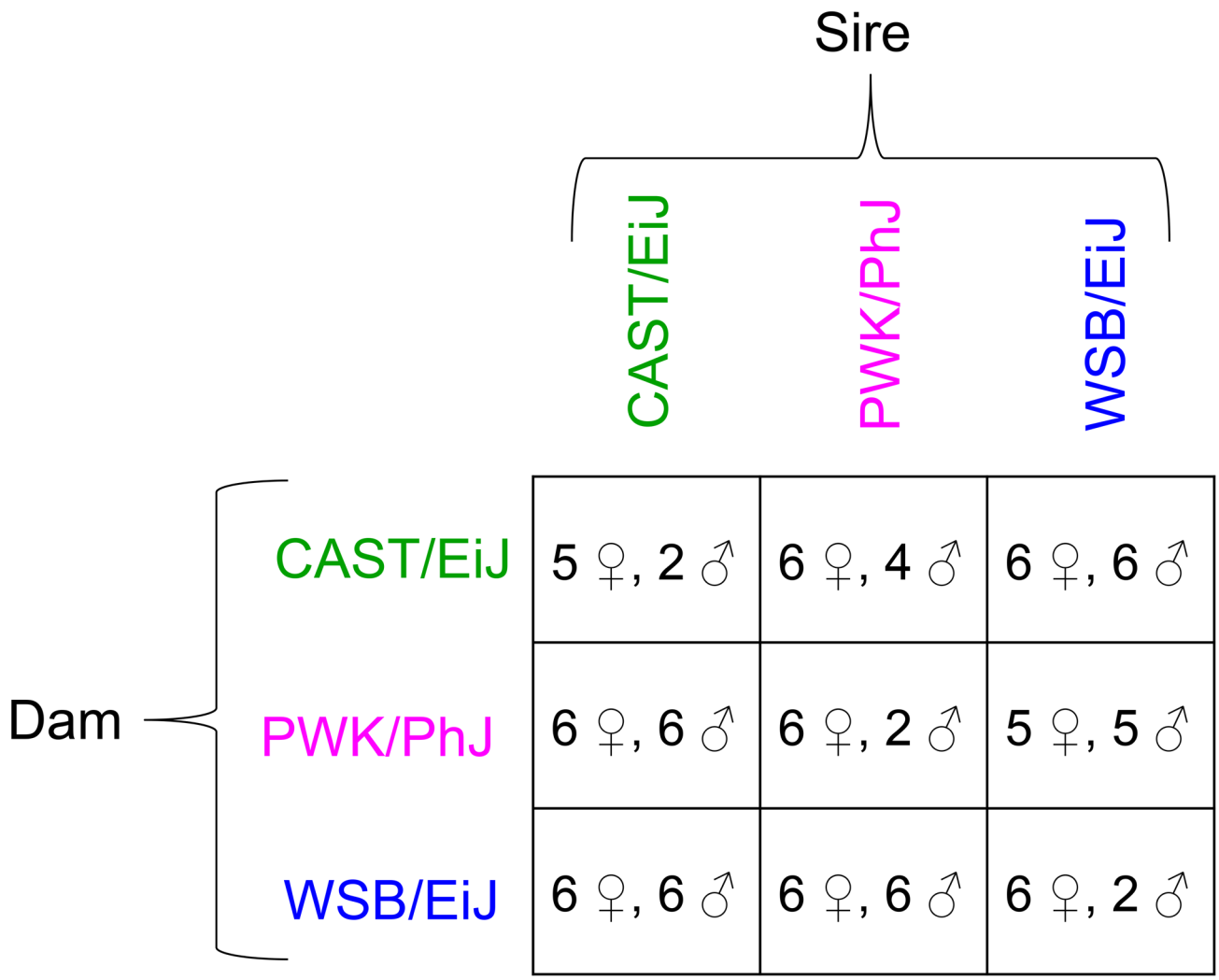

Figure 1.

Diallel crossing scheme and sample sizes. We selected three divergent inbred strains representative of three subspecies within the *Mus 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 days, sacrificed, and total RNA extracted from whole brain, liver, kidney, and lung. Sample size shown is for RNAseq (52 female, 39 male). RNAseq was performed on RNA extracted from brain and microarrays were run on RNA extracted from brain, liver, kidney, and lung.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

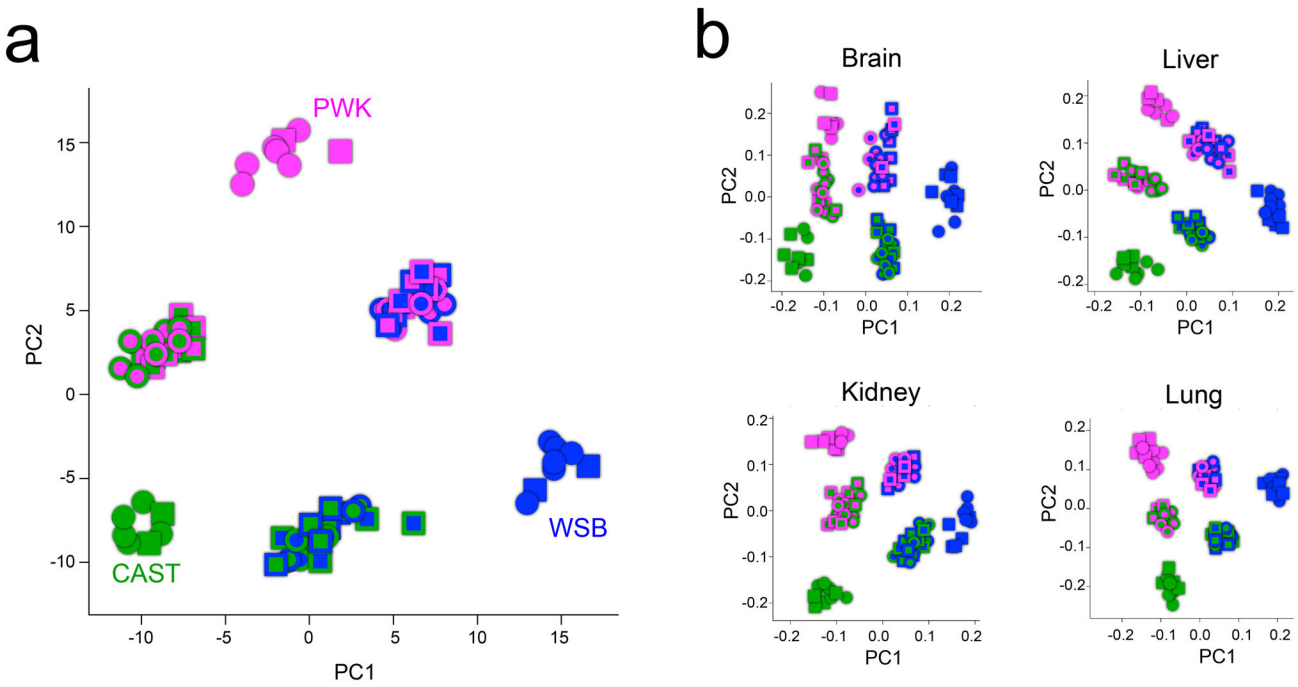

Figure 2.

Principal components of brain RNAseq and microarray expression levels across four tissues. Each point represents one animal with shape indicating sex (circle = female, square = male) and color indicating genotype. For the F_1 animals, the outer color indicates maternal strain and the inner color paternal strain. (**a**) PC1 versus PC2 of the brain RNAseq total read counts 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, greatly exceeding the effects 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 the brain extends to multiple diverse tissues.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

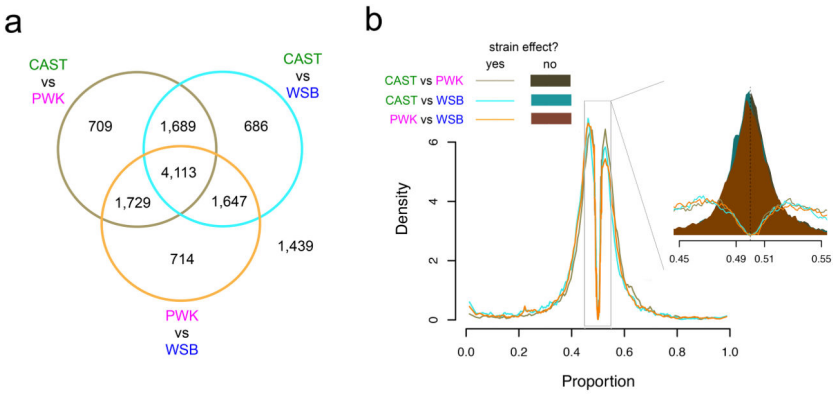

Figure 3.

Balanced contribution of different subspecies to the identification of *cis* regulated genes. (**a**) Venn diagram showing the number of genes with allelic imbalance (FDR < 0.05) in each cross and the relationship to other crosses. (**b**) Distribution of the allelic imbalance effect size for the 11,287 autosomal genes that showed allelic imbalance in 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 or WSB). The inset magnifies the distribution of effect sizes in the vicinity of 0.5 and provides, in the background, the distribution of effect size for genes that did not reach statistical significance for a strain effect (filled distributions).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

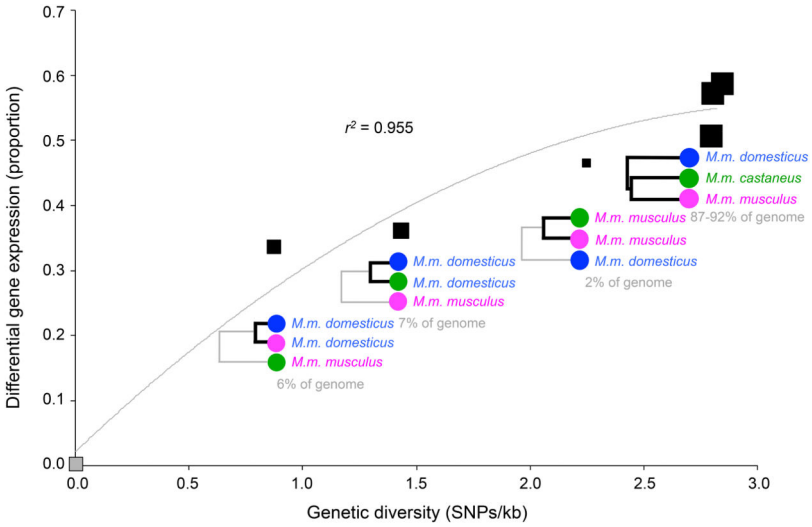

Figure 4.

Differential gene expression is positively correlated to 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, blue: WSB, green: CAST), while colored text represents the subspecific origin in the regions of the genome considered (magenta: *musculus*, blue: *domesticus*, green: *castaneous*). 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 in the figure: 50 Mb, 150 Mb, 175 Mb and 2.25 Gb for the final three comparisons.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

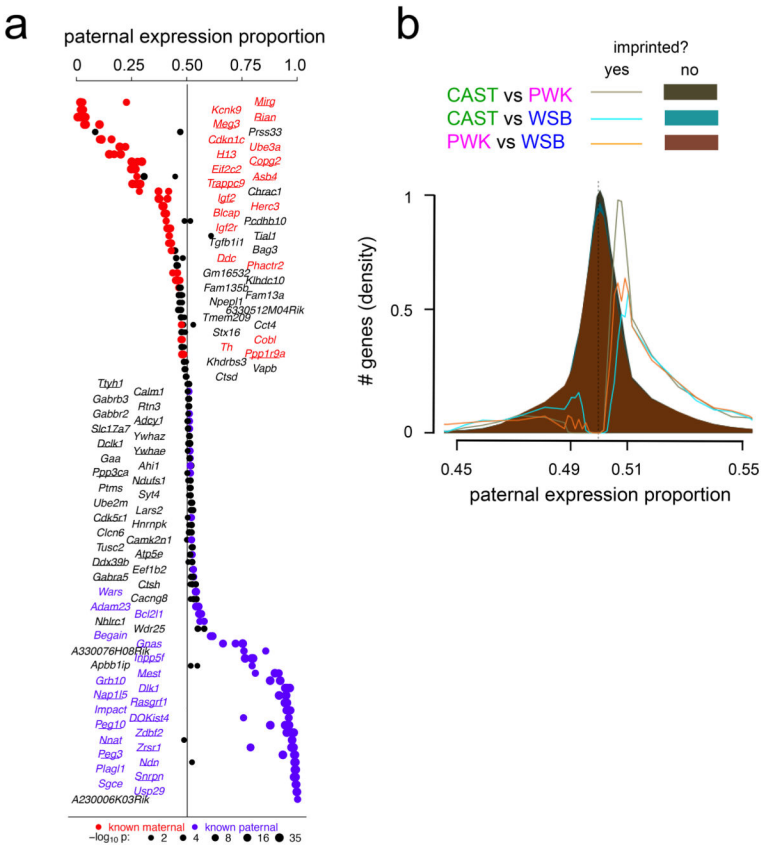

Figure 5.

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 (e.g., CASTxPWK vs PWKxCAST) 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 in blue, and novel imprinted genes in black (n = 54 novel genes). Genes with a strain by 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).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

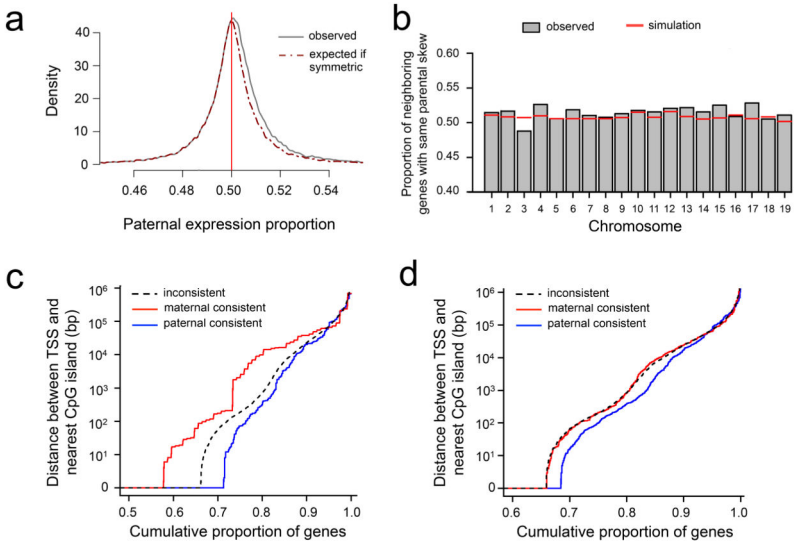

Figure 6.

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 was present). (**b**) Genes with consistent allelic imbalance (found in all three crosses) are clustered in most autosomes. The red lines denote the expected proportion of clustering based on the number of genes with consistent paternal or maternal expression in every autosome. (**c**) Genes with consistent paternal expression in all three crosses and both sexes (N = 467) tend to be closer to CpG islands, while those with consistent maternal expression (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 transcriptional start site (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 expression (N = 3,338) retain enrichment for CpG islands, while those with consistent maternal expression (N = 1,631) are not different from inconsistent genes (N = 5,154).