Unraveling the association between mRNA expressions and mutant phenotypes in a genome-wide assessment of mice

Ben-Yang Liao[1] and Meng-Pin Weng

Division of Biostatistics & Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County 350, Taiwan, Republic of China

Edited by Wen-Hsiung Li, University of Chicago, Chicago, IL, and approved February 25, 2015 (received for review August 6, 2014)

High-throughput gene expression profiling has revealed substantial leaky and extraneous transcription of eukaryotic genes, challenging the perceptions that transcription is strictly regulated and that changes in transcription have phenotypic consequences. To assess the functional implications of mRNA transcription directly, we analyzed mRNA expression data derived from microarrays, RNA-sequencing, and in situ hybridization, together with phenotype data of mouse mutants as a proxy of gene function at the tissue level. The results indicated that despite the presence of widespread ectopic transcription, mRNA expression and mutant phenotypes of mammalian genes or tissues remain associated. The expression-phenotype association at the gene level was particularly strong for tissue-specific genes, and the association could be underestimated due to data insufficiency and incomprehensive phenotyping of mouse mutants; the strength of expression-phenotype association at the tissue level depended on tissue functions. Mutations on genes expressed at higher levels or expressed at earlier embryonic stages more often result in abnormal phenotypes in the tissues where they are expressed. The mRNA expression profiles that have stronger associations with their phenotype profiles tend to be more evolutionarily conserved, indicating that the evolution of transcriptome and the evolution of phenome are coupled. Therefore, mutations resulting in phenotypic aberrations in expressed tissues are more likely to occur in highly transcribed genes, tissue-specific genes, genes expressed during early embryonic stages, or genes with evolutionarily conserved mRNA expression profiles.

mRNA abundance | tissue specificity | developmental stages | ectopic expression | molecular evolution

It is widely assumed that transcription is under tight and sophisticated regulation (1) and that changes in transcription have phenotypic consequences. For example, spatial and temporal regulatory changes in transcription can impart major changes in the development of multicellular organisms (2, 3). Aberrations in transcription have been linked to the onset or progression of human diseases, including autism (4), schizophrenia (5), congenital heart defects (6), and cancers (7). The evolution of transcription regulation was proposed to have a more profound role than protein structural evolution in generating adaptive changes that lead to phenotypic diversity among species (8, 9). However, recent studies exploiting high-throughput gene expression profiling methods suggested widespread ectopic (or nonfunctional) mRNA expression (10) in eukaryotic genomes, according to the discoveries of pervasive transcriptional activities from nongenic regions (11, 12), coexpression in gene clusters without linkage conservation (13) or relatedness in annotated functions (14) of genes, neutral evolution of mRNA expression patterns among orthologs (15), and less correlation in mRNA abundance than in protein abundance among orthologous genes (16). Because the association between mRNA expression and gene function had never been directly assessed systematically and on a genome-wide basis, the general implications in function of mRNA expression became elusive.

The most straightforward approach to discern gene function is mutagenesis. Among mammals, the house mouse (*Mus musculus*) has been subjected to extensive mutagenesis (17), and more than 40% of mouse genes have been mutated, with the resulting mutant strain phenotyped (18). Using these data, we investigate whether genes function in the tissues where they are transcribed. Gene function is defined by the presence of abnormal phenotypes when a gene is mutated. Gene expression data come from oligonucleotide microarray, RNA-sequencing (RNA-seq), and RNA in situ hybridization. The combination of these phenotype data with spatial and temporal mRNA gene expression data (below) allows us to investigate how mRNA expression and phenotypes are connected in the presence of ectopic transcription. In addition, we investigate whether any association in mRNA expression and phenotypes affects the evolutionary conservation of gene expression. Our approach enables us to understand the underlying causes for and the biological features associated with the variations in expression-phenotype association among genes and tissues.

Results and Discussion

Measuring Expression-Phenotype Association. Each gene has an mRNA expression profile, defined as the mRNA expression across the mouse tissues examined. Depending on the experiment (discussed in the following sections), mRNA expression indicates the presence or absence of detectable mRNA expression signals, mRNA expression within a certain range of abundance, or mRNA expression under a specific condition. Each gene also has a phenotypic profile, defined by the presence or absence of abnormal phenotypes in tissues when that gene is mutated. Similarly, each tissue has an mRNA expression profile

Significance

The general perception that mRNA transcription of a gene is regulated by the functional requirements of the cell or tissue has never been systematically tested using genome-wide data. To assess the functional implications of mRNA transcription in mice directly, we analyzed gene expression data and phenotype data of mouse mutants as a proxy of gene function at the tissue level. Our results confirmed the important role transcriptional regulation has in maintaining the gene's proper spatial and temporal functions. In addition, we found that mutations resulting in phenotypic defects in expressed tissues are more likely to occur in highly transcribed genes, tissue-specific genes, genes expressed during early embryonic stages, or genes with evolutionarily conserved mRNA expression.

Author contributions: B.-Y.L. designed research; B.-Y.L. and M.-P.W. performed research; M.-P.W. analyzed data; and B.-Y.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: liaoby@nhri.org.tw.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1415046112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1415046112

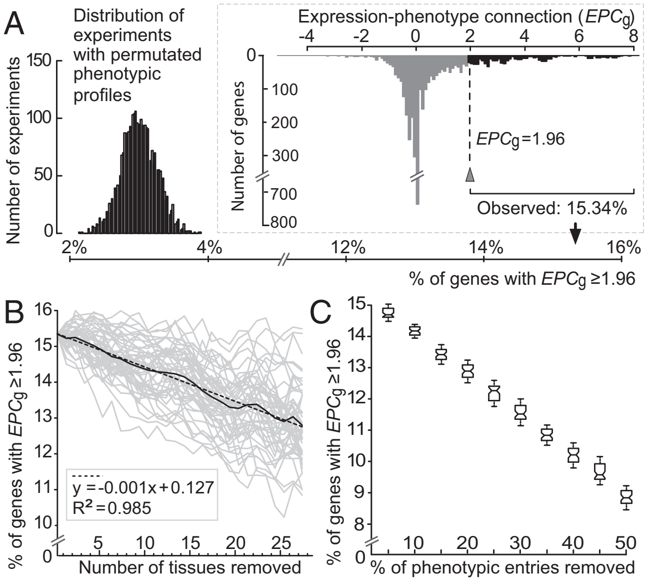


Fig. 1. *EPC* at the gene level, measured as EPC_g. (*A*) Percentage of mouse genes with $EPC_g \geq 1.96$ was 15.34% (arrowhead from the distribution in *Inset*), which was significantly greater than the percentage of mouse genes with $EPC_g \geq 1.96$ estimated from permutation experiments. (*B*) Percentage of mouse genes with $EPC_g \geq 1.96$ decreased linearly with the number of mouse tissues removed in calculating EPC_g. Each gray line is one of 50 experiments derived from the random removal of tissues one at a time. The solid black line is the average of the 50 gray lines. The linear regression line (dashed line with equation given in the gray box) for the average line is given. (*C*) Box plots for the percentage of mouse genes with $EPC_g \geq 1.96$ when a proportion of phenotypic entries were removed. Each box presents the distribution of estimations calculated based on 50 replicates. The values of the upper quartile, median, and lower quartile are indicated in each box, whereas the bars outside the box indicate semiquartile ranges.

defined by mRNA expression across all mouse genes surveyed. For each tissue, there is a phenotypic profile in which there is the presence or absence of abnormal phenotypes for each gene in that given tissue. Using these mRNA expression profiles and phenotypic profiles for mouse genes and tissues (*Methods*), we determined the extent to which mRNA expression profiles correspond to phenotypic profiles for each mouse gene or tissue.

The index for expression-phenotype connection (*EPC*) for each mouse gene or each mouse tissue was defined by the statistical deviation of the observed $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ from the expectation of randomness. When *EPC* is computed for a gene (EPC_{g}), N_{E} is the number of tissues where the gene is expressed, N_{P} is the number of tissues with at least one abnormal phenotype when the gene is mutated, and N_{EP} is the number of tissues that both have gene expression and abnormal phenotypes when the gene is mutated. When *EPC* is computed for a tissue (EPC_{t}), N_{E} is the number of genes that are expressed in the tissue; N_{P} is the number of genes that, when mutated, result in abnormal phenotypes in the tissue; and N_{EP} is the number of genes that both result in mutant phenotypes and exhibit mRNA expression. For each gene or tissue, the distributions of $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ under the null hypothesis of randomness were derived from 2,500 permutation experiments, each of which has a recomputed $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ by randomizing its phenotype profile while maintaining its mRNA expression profile. *EPC* was then defined by the $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ of the original data minus the averaged $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ of the 2,500 permutation experiments divided by the SD of $N_{\mathrm{EP}}/\sqrt{(N_{\mathrm{E}}{\times}N_{\mathrm{P}})}$ derived from 2,500 permutation experiments. We only calculated *EPC* when $N_{\mathrm{E}} \geq 1$ and $N_{\mathrm{P}} \geq 1$. *EPC* is equivalent to the Z-score in the Z-test methodologically. Fig. S1*A* shows the expression and phenotypic profiles of example genes, as well as their corresponding EPC_{g} values.

***EPC* at the Gene Level.** To determine whether mutations in a gene result in defects in tissues where the gene is expressed, we calculated mouse EPC_{g} values using microarray-based mRNA expression data. Microarray expression signals were processed by the gcRMA (GeneChip robust multiarray averaging) method (19), and signals ≥200 indicated that a gene was expressed in a tissue (20) and were used to define N_{E} and N_{EP}. GeneAtlas v2 contains mRNA expression data from oligonucleotide array experiments on 60 mouse tissues ("spinal cord upper" and "spinal cord lower" expression data were merged into "spinal cord" for this study) (20). Of these 60 tissues, 47 have phenotype entries in Mouse Genome Informatics (MGI) from mutant strain phenotyping (*Methods* and Table S1). At present, GeneAtlas v2 contains the largest number of mouse tissues profiled for mRNA expression in a single study, allowing us to measure EPC_{g} with minimal biases (below). In the distribution of EPC_{g} from 3,859 mouse genes with $N_{\mathrm{E}} \geq 1$ and $N_{\mathrm{P}} \geq 1$ in 47 tissues (Fig. 1*A*), 15.34% (592 of 3,859) of the genes have $EPC_{\mathrm{g}} \geq 1.96$ ($P < 0.05$, Z-test), which is significantly greater than the percentage of genes with $EPC_{\mathrm{g}} \geq 1.96$ in the permutation experiments (3.00 ± 0.26% SD; $P < 10^{-300}$, *t* test; Fig. 1*A*). For genes above this EPC_{g} threshold, mRNA expression signals are directly tied to function in the tissues where they are expressed, and the loss of this function can result in abnormal phenotypes. A larger EPC_{g} threshold resulted in a greater deviation of observed EPC_{g} from the expectation under randomness (Fig. S2). Focusing on 1,216 tissue-specific genes ($N_{\mathrm{E}} \leq 5$), the proportion of genes with $EPC_{\mathrm{g}} \geq 1.96$ is 36.51% (444 of 1,216 genes). Therefore, the expression-phenotype association was observed despite intrinsic noise in mRNA expression (21) and measurement errors of microarrays (22) (below), and the association was particularly strong for tissue-specific genes.

Although significantly higher than random, the percentage of mouse genes with statistically significant *EPC* (15.34% of all genes) was small. For the remaining genes, there are two possible explanations for a lack of statistical support for *EPC*. First, a mutation in a gene might not cause discernable defects in the tissue where it is expressed. Second, there is a connection between mRNA expression and mutant phenotypes for a gene, but it cannot be detected due to data limitations, such as insufficient tissue sampling, incomplete phenotyping, and errors in mRNA quantification. A mouse has >150 cell types (23), which comprise even more tissues, but only 47 tissues were included in this analysis (Fig. 1*A*). To understand the influence of incomplete tissue sampling in estimating EPC_{g}, we randomly removed the data on a tissue one at a time until only 20 tissues remained and recalculated EPC_{g} after each random tissue removal. Based on 50 replicates of random tissue exclusion, we averaged the percentage of genes with $EPC_{\mathrm{g}} \geq 1.96$ for each number of tissues removed (Fig. 1*B*). If the percentage was not affected by incomplete tissue sampling, the percentage should plateau to a value of ∼15.34% when the number of tissues removed is few; after a certain stage, the percentage decreases as the number of tissues removed increases. Alternatively, if the percentage was underestimated due to insufficient tissue sampling, the percentage of genes with $EPC_{\mathrm{g}} \geq 1.96$ should decrease as the number of tissues removed increases from the beginning. Consistent with the insufficient tissue sampling, we found that the percentage of genes with $EPC_{\mathrm{g}} \geq 1.96$ decreases linearly as the number of tissues removed increases ($R^2 = 0.985$; $P < 10^{-300}$, ANOVA for linear model fits; Fig. 1*B*).

Because phenotype screening of mutant mice can be biased by study design, the manifestation of a mutation in tissues unrelated to the study focus can be overlooked and remain undescribed. Thus, although 47 tissues are included in our analysis, for most genes, only a fraction of these tissues were examined for phenotypic abnormalities. To account for incomplete phenotyping, we randomly removed a proportion of phenotypic entries

(5–50% in 5% increments) and recalculated the percentage of mouse genes with $EPC_g \geq 1.96$. Each of these 10 experiments had 50 replicates. The median percentage of genes with $EPC_g \geq 1.96$ decreased from 14.70% to 8.83% as the proportion of phenotypic entries removed increased (Fig. 1*C*), indicating that incomplete phenotyping leads to an underestimation of EPC_g. When expression and phenotype data are available for more tissues (Fig. 1*B*) and when phenotyping of mutant strains is more complete (Fig. 1*C*), the percentage of genes with $EPC_g \geq 1.96$ will likely increase toward its true value.

Phenotyping mutant mouse strains can be biased by prior knowledge of a gene's expression pattern. To determine if the significant EPC_g in Fig. 1*A* is due to this bias, we limited the analysis to phenotype data published before the microarray dataset GeneAtlas v2 (20). If high EPC_g values in the full dataset are due to phenotype inspection bias, EPC_g of this subset of 2,084 genes should be lower. However, 16.21% (338 of 2,084) of phenotyped mouse genes had $EPC_g \geq 1.96$ (Fig. S3), which was slightly greater but not statistically different from the full dataset ($P = 0.37$, χ^2 test). Therefore, the observed EPC_g values were not due to biased phenotypic screening of tissues/organs with known gene expression signals.

Microarray signals harbor cross-hybridization noises (22). To account for experimental errors in the estimation of EPC_g, we stochastically introduced noise within a range (±5% to ±50% of the experimental value) into the microarray signals and recalculated EPC_g. Regardless of the magnitude of noise introduced, the median percentage of mouse genes with $EPC_g \geq 1.96$ stayed near 15.4% (Fig. S4), indicating that microarray noise had no effect on the estimate of EPC_g.

Genes Without Overlapping Tissues of mRNA Expression and Mutant Phenotypes. Of 3,859 genes in the full dataset, 996 genes resulted in abnormal phenotypes in tissues where they were not expressed (Fig. 1*A*; genes with $N_{EP} = 0$, $N_E > 0$, $N_P > 0$). These genes had an average EPC_g of −0.451 ± 0.338 SD, which was smaller than the average EPC_g of the rest of the genes (1.064 ± 1.748 SD). Incomplete phenotyping and insufficient tissue sampling that underestimated EPC_g (Fig. 1 *B* and *C*) can explain part of $N_{EP} = 0$ genes. For example, although beta-1,4-N-acetyl-galactosaminyl transferase 1 gene (*B4galnt1*) is expressed in many adult tissues (Fig. S1*B*), only reproductive and neurological phenotypes have been examined for the viable and fertile knockout strain (24, 25). More complete phenotyping on *B4galnt1* mutant strains across a wider range of tissues could reveal overlap between tissues with expression and tissues with abnormalities. All 47 tissues used to compute EPC_g were from adult mice, but many abnormal phenotypes are only observed during fetal or neonatal stages due to premature death. For example, embryos from an NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase gene (*Mthfd2*) knockout strain have small pale livers and die before embryonic day 15.5, suggesting an important role of *Mthfd2* in embryogenesis (26). Although the expression dataset indicates that *Mthfd2* is highly expressed in the fertilized egg and embryonic stages (Fig. S1*C*), this information was not used in the calculation of EPC_g. Another explanation for $N_{EP} = 0$ genes is a bias in abnormal phenotypes that affect the whole organism rather than a specific tissue or cell type. Using Mammalian Phenotype Enrichment Analysis (MamPhEA) (18) for phenotypic enrichment analyses (*Methods*), compared with other genes ($N_{EP} > 0$), $N_{EP} = 0$ genes were enriched in phenotypes associated with the nervous system, behaviors, and cellular metabolism/homeostasis (the full results are shown in Fig. S5). These phenotypes have organism-level effects (e.g., behavioral changes, obesity) that cannot simply be coded as an abnormality in specific tissues. When these abnormalities are observed in tissues, they are often due to gross physiological changes in the mutant (e.g., increased/decreased fat amount due to changes in metabolic rate) rather than dysfunction of locally expressed genes. These examples indicated that the importance of mRNA transcription to gene function could be underestimated by the data and method used in this study.

***EPC* at the Tissue Level.** To assess the functional relevance of the transcriptome at the tissue level to that tissue's morphology or physiology directly, we examined EPC_t of 47 mouse tissues for 7,449 genes that have MGI phenotype entries and have mRNA expression profiles in GeneAtlas v2. Each mouse tissue has two gene profiles: one including genes with active (expression signals ≥200) transcription and another including genes that, when mutated, result in abnormal phenotypes in the tissue examined. To compute EPC_t, the first expression profile was used to define N_E and the second phenotypic profile was used to define N_P. The two profiles were used together to define $N_{EP.}$ Of the 47 tissues, 22 (46.81%) have statistically significant *EPC*, indicated by $EPC_t \geq 1.96$ (Fig. 2*A*). By comparison, the average percentage of

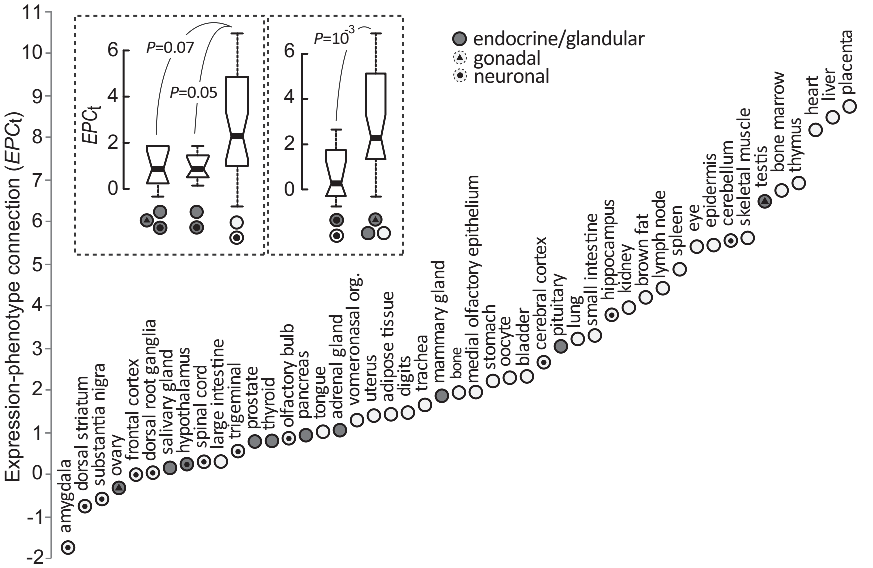


Fig. 2. *EPC* at the tissue level, measured by EPC_t, in 47 mouse tissues. Comparisons of EPC_t between glandular/endocrine tissues (with and without gonadal tissues) vs. nonglandular/nonendocrine tissues or between neuronal tissues vs. nonneuronal tissues are shown (P values are from the Mann–Whitney U test). The values of the upper quartile, median, and lower quartile are indicated in each box, whereas the bars outside the box indicate semiquartile ranges. org., organ.

tissues that had $EPC_t \geq 1.96$ from permuted phenotypic profiles of each of the tissues was 2.52% (Fig. S6). A complementary approach exploiting hypergeometric tests (18) echoed the analysis by EPC_t, showing that only gene sets expressed in tissues with high EPC_t have significant statistical support for the enrichment of abnormal phenotypes in the same tissue (Table S2). These results suggest that genes transcribed in a tissue are often performing functions linked to the development and physiology of the tissue examined.

Tissue function likely underlies the observed variation in EPC_t (Fig. 2). If colocalization of transcription and protein activity (either directly or indirectly) leads to high EPC_t, tissues with primarily endocrine or glandular functions, which produce molecules that control the development or physiology of other tissues or interact with external environmental factors, should have smaller EPC_t. Mutations in genes transcribed in these tissues manifest themselves as abnormal phenotypes in other tissues. The 47 mouse tissues include 10 endocrine/glandular tissues (Fig. 2). As predicted, these 10 endocrine/glandular tissues had lower EPC_t compared with the remaining 37 tissues, although the difference was not statistically significant ($P = 0.07$, Mann–Whitney U test; left box in Fig. 2, *Top Left*). The 10 endocrine/glandular tissues include testis and ovary, which not only secrete hormones that function outside the tissue, but are involved in reproductive activities, such as the production of gametes. When testis and ovary were excluded from the 10 endocrine/glandular tissues, the difference in EPC_t between the eight endocrine/glandular tissues and the 37 nonendocrine/glandular tissues was significant ($P < 0.05$, Mann–Whitney U test; Fig. 2). The higher EPC_t of tissues producing proteins/molecules that function locally suggests that the location of transcription has functional implications.

Additionally, genes associated with neurological/behavioral traits [e.g., VGF nerve growth factor inducible (*Vgf*); Fig. S1*A*] can have global effects, which lead to lack of observable EPC (also Fig. S5). Consistent with this observation, neuronal tissues also tend to have lower EPC_t compared with other tissues ($P = 0.001$, Mann–Whitney U test; right box in Fig. 2, *Top Left*).

mRNA Abundance and *EPC*. mRNA abundance can vary by several orders of magnitude in mammalian cells. Highly expressed genes are more easily identified experimentally (e.g., cDNA cloning) and tend to be well studied. To understand how mRNA abundance of a gene relates to its functional relevance, we used RNA-seq data from five mouse tissues (cerebellum, heart, kidney, liver, and testis) profiled in a study by Brawand et al. (27). These same investigators also profiled the tissue "brain," but we omitted this organ because there was no corresponding "whole-brain" phenotypic code in MGI. RNA-seq data were used because RNA-seq quantifies mRNA abundance more accurately compared with microarrays and has low background noise to detect weakly expressed genes (28). RNA-seq expression signals were measured as reads per kilobase per million mapped reads (RPKM) (27). The number of transcribed genes (RPKM > 0) in cerebellum, heart, kidney, liver, and testis was 4,389, 4,326, 4,317, 4,022, and 4467, respectively. For each tissue, we divided all transcribed genes into five equal-sized bins based on mRNA abundance (1–5, lowest to highest). For each bin of each tissue, N_E and N_{EP} were defined by gene counts. For example, in bin 2 of the heart, N_E was defined by the number of genes transcribed in the heart bin 2 (at 20–40% abundance rank), and the number of these genes also found to have abnormal phenotypes in the heart was N_{EP}. We calculated the proportion of expressed genes showing abnormal phenotypes (N_{EP}/N_E) and EPC_t for each bin of each tissue. When computing EPC_t, 7,358 phenotyped genes with detectable RNA-seq expression signals (RPKM > 0) in at least one of the five tissues were used for the permutation analysis (2,500 permutations).

Because protein production can be regulated posttranscriptionally and proteins vary in their intrinsic range of functionally effective abundance levels (29), highly transcribed genes in the cell are not necessarily functionally more important than lowly expressed genes. However, the proportion of genes expressed resulting in abnormal phenotypes (N_{EP}/N_E) increased as the bin number increased for all of the tissues (Fig. 3*A*), indicating that genes with higher mRNA expression are more likely to be associated with mutant phenotypes. For bin 1 of all tissues (and also for bin 2 for testis and bins 2 and 3 for cerebellum), EPC_t was <1.96 (Fig. 3*B*), suggesting lowly expressed genes are largely composed of genes functionally unimportant to the tissues where they are expressed. These results correspond to two major classes of transcribed genes with distinct mRNA abundances in metazoan cells: highly expressed, functionally important genes and lowly expressed genes with ectopic expression (30). N_{EP}/N_E and EPC_t increased with mRNA abundance in all of the tissues, even among the highly expressed genes (i.e., bins 3–5), indicating that even among highly expressed genes, the absolute difference in mRNA abundance in tissues reflects a difference in the gene's importance to tissue function.

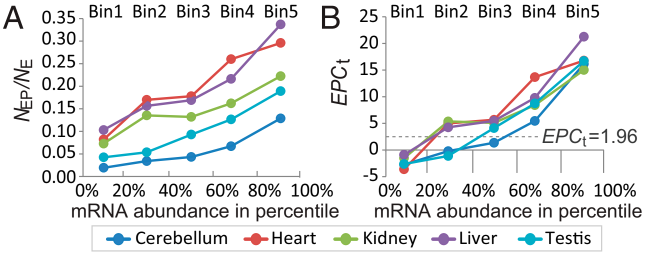


Fig. 3. Proportion of genes that causes phenotypic abnormalities in the tissue where expressed (N_{EP}/N_E) (*A*) and EPC_t values (*B*) were higher for genes that have higher expression levels, defined by RNA-seq expression signals presented as RPKM. Genes were binned into five equal-sized bins based on mRNA abundance (1–5, lowest to highest) for five tissues.

Expression in Embryonic Stages and *EPC*. Analysis of the spatial and temporal regulation of genes during embryogenesis is necessary to understand developmental genes. In situ hybridization is a classic technique used to visualize mRNA expression in embryos, although in situ hybridization expression signals are noisy and detection specificity is affected by probe design (31). Changes in gene regulation at earlier stages in embryogenesis tend to result in more severe phenotypes, indicated by a higher frequency of embryonic lethality (32), and enhancers responsible for earlier stages of organogenesis are proposed to be more evolutionarily conserved (33), suggesting that early regulatory activities are especially important. Therefore, temporal expression during embryogenesis could be a factor determining the influences of a gene's expression to the physiology or morphology of the tissue.

To test the hypothesis, we investigated two tissues: eye and heart. These tissues have discernable primordia with distinct developmental processes from their surrounding tissues, and their development is observable in early embryonic stages. mRNA in situ hybridization data from mouse embryos was retrieved from the e-Mouse Atlas of Gene Expression (EMAGE) (34) (*Methods*), which integrates in situ hybridization expression data from diverse sources. In the EMAGE data, Edinburgh Mouse Atlas Project (EMAP) IDs differ for embryonic tissues at different developmental Theiler stages. For eye and heart, expression data for Theiler stages 12–23 were included (Table S3). If a gene is expressed at multiple stages, the earliest stage was assigned for that gene. The number of genes with in situ hybridization signals in the developing eye (or heart) at each Theiler stage was defined as N_E, and the number of genes that also showed mutant phenotypes in the adult eye (or heart) was defined as N_{EP}. Stage-specific N_{EP}/N_E and EPC_t were only computed for Theiler

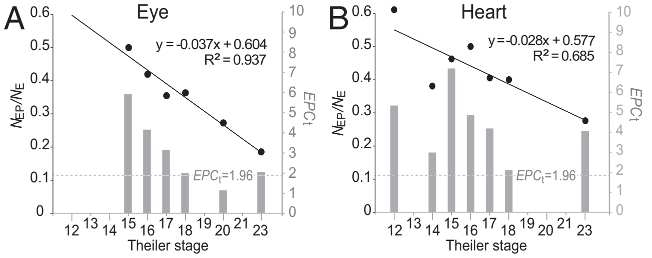


Fig. 4. Proportion of genes that causes phenotypic abnormalities in the tissue where expressed (N_{EP}/N_E) (black dots correspond to the left *y* axis) and EPC_t values (gray bars correspond to the right *y* axis) for different embryonic stages of mouse developing eye (*A*) or heart (*B*). The linear regression line (and the equation) for N_{EP}/N_E and Theiler stages are shown.

stages with a sufficient sample size ($N_E \geq 10$). When computing EPC_t, 3,043 genes that have MGI phenotype data and EMAGE in situ hybridization data were used for phenotypic profile permutation (2,500 permutations). For both eye and heart, genes expressed at earlier Theiler stages in primordial embryonic tissues are more likely to result in phenotypic defects in the corresponding adult tissue (Theiler stages vs. N_{EP}/N_E: eye: Spearman's correlation coefficient σ = –0.943, $P < 10^{-3}$; heart: σ = –0.824, $P = 0.02$; Fig. 4*A*). This conclusion is also supported by EPC_ts, which are >1.96 for genes expressed at early embryonic stages (before Theiler stage 18) of both eye and heart (Fig. 4*B*).

***EPC* and the Evolution of Gene Regulation.** Natural selection acts on phenotypes that have an effect on an organism's fitness. Gene properties tightly connected with phenotypes thus should more often be the subject of natural selection. To examine whether genes with a higher EPC_g have more evolutionarily constrained mRNA expression profiles, we computed expression profile divergence between 1:1 human-mouse orthologs. Expression profile divergence was measured by $1 - R$, where R is Pearson's correlation coefficient between microarray expression signals across the 26 homologous tissues between human and mouse in GeneAtlas v2 (22) (*Methods*). Higher $1 - R$ indicates a greater expression profile divergence and more relaxed selective constraint in expression profile. Consistent with our expectation, $1 - R$ of genes with $EPC_g > 1.96$ was significantly lower than $1 - R$ of other genes (Fig. 5*A*). This difference was not observed under the neutral model of transcriptome evolution approximated by randomizing expression profiles of human genes in calculating $1 - R$ (22, 35) (Fig. S7). Hence, mRNA expression profiles that are associated with the phenotypic profiles tend to be more evolutionarily conserved after the rodent-primate divergence.

Evolutionarily conserved expression profiles, indicated by high $1 - R$, were found in tissue-specific genes in mammals (36). We found that genes with high EPC_g tend to be genes with high tissue specificity, indicated by lower N_E or higher τ (*Methods*) (EPC_g vs. N_E: σ = –0.081, $P < 10^{-4}$; EPC_g vs. τ: σ = 0.163, $P < 10^{-14}$), suggesting that tissue specificity has potentially confounded the relationship between EPC_g and $1 - R$. To measure the direct association between EPC_g and $1 - R$, we performed partial correlation analysis. In addition to τ, N_E, and N_P, we examined and controlled for other gene properties potentially governing regulatory evolution of genes, including gene essentiality (*Essen*), microarray-based mRNA abundance (*ExpAb*), number of associated Gene Ontology (GO) terms (GO_M, GO_B, or GO_C represents the number of terms in "molecular function," "biological processes," or "cellular components," respectively; *Methods*) and number of interacting partners in the protein–protein interaction network (K_{PPI}). The partial rank correlation coefficient (σ_p) between $1 - R$ and EPC_g after controlling for all of the other factors remained significantly negative (σ_p = –0.108, $P < 10^{-6}$) (Fig. 5*B*). Thus, genes with evolutionarily constrained mRNA expression profiles tend to be associated with abnormal phenotypes in the expressed tissues. Consistent with the previous notion that tissue-specific genes tend to have greater *EPC*s, there was a positive correlation between EPC_g vs. τ and a negative correlation between EPC_g vs. N_E after controlling for all of the remaining factors (Fig. 5*B*).

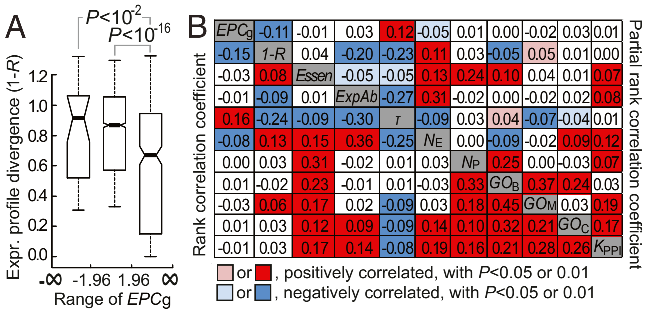
EPC_g is an approximation of how well the mRNA expression pattern of a gene matches its physiological function, because ectopic expression in off-target tissues and a lack of detectable expression in tissue requiring a gene's product decrease EPC_g. The observed pattern that gene expression profiles with higher EPC_g are more evolutionarily conserved supports the argument that gene expression profiles are shaped by purifying selection (22, 37). In addition to previously reported gene properties, such as tissue specificity and mRNA expression level (36), we found that EPC_g is a factor correlated with the rate of regulatory evolution of mammalian genes.

Summary

Despite the presence of widespread ectopic transcription, mRNA expression and mutant phenotypes remain tightly associated at both the gene level and tissue level in mice. The expression-phenotype association at the gene level, indicated by EPC_g, was particularly strong for tissue-specific genes, and could be underestimated in the present study due to incomplete tissue sampling and phenotyping. The variation in the association at tissue level, EPC_t, can be explained by tissue functions. Mutations on genes expressed at higher levels or expressed at earlier embryonic stages more often result in abnormal phenotypes in the tissues where they are expressed. The mRNA expression profiles that have stronger connections with their phenotype profiles tend to be more evolutionarily conserved, indicating that the evolution of transcriptome and phenome are coupled. Our results suggest that changes in mRNA expression that cause more severe abnormal phenotypes or diseases in expressed tissues are more likely to occur in genes with abundant transcription levels, high tissue specificities, early expressed embryonic stages, or evolutionarily conserved expression profiles.

Methods

Phenotype Data of Mouse Genes. Mouse genes and the associated mutant phenotypes were obtained from MGI (www.informatics.jax.org/), version 5.2. Ensembl IDs (v69) (38) of phenotyped mouse protein-coding genes



EPC_g	-0.11	-0.01	0.03	0.12	-0.05	0.01	0.00	-0.02	0.03	0.01
-0.15	*1-R*	0.04	-0.20	-0.23	0.11	0.03	-0.05	0.05	0.01	0.00
-0.03	0.08	*Essen*	-0.05	-0.05	0.13	0.24	0.10	0.04	0.01	0.07
-0.01	-0.09	0.01	*ExpAb*	-0.27	0.31	-0.02	0.00	-0.02	0.02	0.08
0.16	-0.24	-0.09	-0.30	τ	-0.09	0.03	0.04	-0.07	-0.04	0.01
-0.08	0.13	0.15	0.36	-0.25	N_E	0.00	-0.09	-0.02	0.09	0.12
0.00	0.03	0.31	-0.02	0.01	0.03	N_P	0.25	0.00	-0.03	0.07
0.01	-0.02	0.23	-0.01	0.01	-0.03	0.33	GO_B	0.37	0.24	0.03
-0.03	0.06	0.17	0.02	-0.09	0.03	0.18	0.45	GO_M	0.03	0.19
0.01	0.03	0.12	0.09	-0.09	0.14	0.10	0.32	0.21	GO_C	0.17
-0.01	0.03	0.17	0.14	-0.08	0.19	0.16	0.21	0.28	0.26	K_{PPI}

▢ or ■, positively correlated, with P<0.05 or 0.01
▢ or ■, negatively correlated, with P<0.05 or 0.01

Fig. 5. Genes with higher EPC_g have more conserved gene expression profiles following the rodent-primate divergence. (*A*) Box plots show the distributions of $1 - R$ for mouse genes with three different ranges of EPC_g. The values of the upper quartile, median, and lower quartile are indicated in each box, and the bars outside the box indicate semiquartile ranges. *P* values are from the Mann–Whitney *U* test. (*B*) Rank correlation between EPC_g and $1 - R$ before and after controlling for potential confounding factors. Rank correlation coefficients between the two examined properties are given in the bottom left corner, whereas the partial rank correlation coefficients after controlling for the all of the remaining factors are given in the upper right corner.

were found at MRK_ENSEMBL.rpt, whereas the information on genotypes and phenotypes [presented as mammalian phenotype IDs (MP IDs)] (39) of the mutant generated was found at MGI_GenePheno.rpt. Phenotypes caused by mutations on multiple genes were discarded. We obtained a dataset of 7,449 mouse genes with one or more MP IDs when the gene is knocked out, knocked down, or mutated by transgenic insertions or point mutations. Organs or other anatomical parts with abnormal phenotypes are specified by MP IDs that are hierarchically structured. A parent MP ID represents a phenotype lineage that may include several child MP IDs to describe a more detailed abnormal phenotype. Genes with a child MP ID were also assigned to the parent MP IDs. MP ID terms used to define abnormal phenotypes in the 47 tissues are listed in Dataset S1. Publication dates for the literature describing abnormal phenotypes for MGI annotations were obtained from PubMed IDs (www.ncbi.nlm.nih.gov/pubmed).

mRNA Expression Data of Mouse Genes. Expression signals of mouse genes measured by mRNA hybridization from 61 mouse tissues to the Affymetrix microarray chip (GNF1M) were obtained from the GeneAtlas v2 dataset (20). The processed mRNA expression signals, calculated by RPKM, for each mouse gene derived from 76-mer RNA-seq experiments in mouse cerebellum, heart, kidney, liver, and testis were obtained from supplementary material of a study by Brawand et al. (27). The mRNA in situ hybridization data from mouse embryonic tissues at various Theiler stages (40) were obtained from the BioMart interface of the EMAGE database (41). Genes annotated with "detected," "strong," "moderate," or "weak" hybridization intensity were considered to be "expressed" in tissues/organs identified by EMAP IDs, according to anatomical ontology of developing mouse embryos (42). EMAP IDs are also hierarchically structured and are Theiler stage-specific. Accordingly, genes with a child EMAP ID were also assigned with the parent EMAP ID. To understand the influence of expression in embryonic stages to the *EPC*, we focused on EMAP IDs associated with eye and heart from Theiler stages 12–23 (Table S3).

Expression Profile Conservation of Mouse Genes. Orthology relationships between human and mouse genes based on Ensembl annotation v69 were retrieved using BioMart (www.biomart.org/). Only 1:1 orthologs were used to compute expression profile divergence, as measured by $1 - R$. The 26 homologous tissues between mouse and human used to compute $1 - R$ are from Liao and Zhang (22). *ExpAb* is defined as the average microarray signal across the 26 tissues. Tissue specificity of a mouse gene is calculated by $\tau = \left[\sum_{j=1}^{n} \left(1 - \left[\frac{\log_2 S(j)}{\log_2 S_{\max}}\right]\right)\right] \Big/ (n-1)$ (36), where $n = 26$ and $S_{\max}$ is the highest expression signal of the gene across the 26 tissues. Following Liao and Zhang (36), we arbitrarily let $S(j)$ be 100 if it is lower than 100. The τ value ranges from 0 to 1, with higher values indicating higher tissue specificity. A gene is essential (*Essen* = 1) when mutations on it lead to premature death or infertility; otherwise, it is nonessential (*Essen* = 0) (43). GO_M, GO_B, or GO_C was defined by the number of associated GO terms at level 5 annotated by Ensembl. The mouse protein–protein interaction network was obtained from the mammalian protein–protein interaction database at Munich Information Center for Protein Sequences (44). To identify factors contributing to $1 - R$, partial correlation analysis was conducted using the "ppcor" package (45) for R (www.r-project.org/).

ACKNOWLEDGMENTS. We thank anonymous reviewers for valuable comments and Wendy Grus for assistance in editing the manuscript. This work was supported by intramural funding from the National Health Research Institutes, Taiwan, and a research grant (MOST 101-2311-B-400-001-MY3) from the Ministry of Science and Technology, Taiwan (to B.-Y.L.).

1. Emerson BM (2002) Specificity of gene regulation. *Cell* 109(3):267–270.
2. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6–21.
3. Forrest ARR, et al.; FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507(7493):462–470.
4. Voineagu I, et al. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474(7351):380–384.
5. Sanders AR, et al.; MGS (2013) Transcriptome study of differential expression in schizophrenia. *Hum Mol Genet* 22(24):5001–5014.
6. Pierpont ME, et al.; American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young (2007) Genetic basis for congenital heart defects: Current knowledge. *Circulation* 115(23):3015–3038.
7. Slamon DJ, deKernion JB, Verma IM, Cline MJ (1984) Expression of cellular oncogenes in human malignancies. *Science* 224(4646):256–262.
8. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188(4184):107–116.
9. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8(3):206–216.
10. Rodríguez-Trelles F, Tarrío R, Ayala FJ (2005) Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? *BioEssays* 27(6):592–601.
11. Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 21(2):93–102.
12. Clark MB, et al. (2011) The reality of pervasive transcription. *PLoS Biol* 9(7):e1000625; discussion e1001102.
13. Liao B-Y, Zhang J (2008) Coexpression of linked genes in Mammalian genomes is generally disadvantageous. *Mol Biol Evol* 25(8):1555–1565.
14. Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol* 1(1):5.
15. Khaitovich P, et al. (2004) A neutral model of transcriptome evolution. *PLoS Biol* 2(5):E132.
16. Schrimpf SP, et al. (2009) Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol* 7(3):e48.
17. Gondo Y (2008) Trends in large-scale mouse mutagenesis: From genetics to functional genomics. *Nat Rev Genet* 9(10):803–810.
18. Weng M-P, Liao B-Y (2010) MamPhEA: A web tool for mammalian phenotype enrichment analysis. *Bioinformatics* 26(17):2212–2213.
19. Wu ZJ, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. *J Am Stat Assoc* 99(468):909–917.
20. Su AI, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101(16):6062–6067.
21. Blake WJ, KAErn M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422(6932):633–637.
22. Liao B-Y, Zhang J (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol Biol Evol* 23(3):530–540.
23. Vogel C, Chothia C (2006) Protein family expansions and biological complexity. *PLOS Comput Biol* 2(5):e48.
24. Liu Y, et al. (1999) A genetic model of substrate deprivation therapy for a glycosphingolipid storage disorder. *J Clin Invest* 103(4):497–505.
25. Takamiya K, et al. (1996) Mice with disrupted GM2/GD2 synthase gene lack complex gangliosides but exhibit only subtle defects in their nervous system. *Proc Natl Acad Sci USA* 93(20):10662–10667.
26. Di Pietro E, Sirois J, Tremblay ML, MacKenzie RE (2002) Mitochondrial NAD-dependent methylenetetrahydrofolate dehydrogenase-methenyltetrahydrofolate cyclohydrolase is essential for embryonic development. *Mol Cell Biol* 22(12):4158–4166.
27. Brawand D, et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–348.
28. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57–63.
29. Heo M, Maslov S, Shakhnovich E (2011) Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proc Natl Acad Sci USA* 108(10):4258–4263.
30. Hebenstreit D, et al. (2011) RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol Syst Biol* 7:497.
31. Arvey A, et al. (2010) Minimizing off-target signals in RNA fluorescent in situ hybridization. *Nucleic Acids Res* 38(10):e115.
32. Damjanovski S, Sachs LM, Shi YB (2000) Multiple stage-dependent roles for histone deacetylases during amphibian embryogenesis: Implications for the involvement of extracellular matrix remodeling. *Int J Dev Biol* 44(7):769–776.
33. Liao BY, Weng MP (2012) Natural selection drives rapid evolution of mouse embryonic heart enhancers. *BMC Syst Biol* 6(Suppl 2):S1.
34. Christiansen JH, et al. (2006) EMAGE: A spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res* 34(Database issue):D637–D641.
35. Jordan IK, Mariño-Ramírez L, Koonin EV (2005) Evolutionary significance of gene expression divergence. *Gene* 345(1):119–126.
36. Liao B-Y, Zhang J (2006) Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Mol Biol Evol* 23(6):1119–1128.
37. Denver DR, et al. (2005) The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nat Genet* 37(5):544–548.
38. Hubbard TJ, et al. (2007) Ensembl 2007. *Nucleic Acids Res* 35(Database issue): D610–D617.
39. Smith CL, Eppig JT (2012) The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mamm Genome* 23(9-10): 653–668.
40. Theiler K (1972) *The House Mouse: Atlas of Embryonic Development* (Springer, New York).
41. Stevenson P, Richardson L, Venkataraman S, Yang Y, Baldock R (2011) The BioMart interface to the eMouseAtlas gene expression database EMAGE. *Database (Oxford)* 2011:bar029.
42. Hayamizu TF, et al. (2013) EMAP/EMAPA ontology of mouse developmental anatomy: 2013 update. *J Biomed Semantics* 4(1):15.
43. Liao B-Y, Zhang J (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet* 23(8):378–381.
44. Pagel P, et al. (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21(6):832–834.
45. Kim SH, Yi SV (2007) Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica* 131(2):151–156.