EXPRESSION EVOLUTION OF MAMMALIAN GENES

by

Ben-Yang Liao

A dissertation submitted in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2008

Doctoral Committee:

Associate Professor Jianzhi Zhang, Chair
Professor Priscilla K. Tucker
Assistant Professor Zhaohui Qin
Assistant Professor Patricia Jean Wittkopp

ACKNOWLEGMENTS

My first thanks go to Jianzhi George Zhang, an exceptional advisor I am lucky enough to have worked with. His kindness, encouragement, and support made all the difference in my graduate studies. His consideration enabled me to balance my study and my family. I am grateful for his support when I had to leave US for months during my second and third years due to family needs. George is amazingly knowledgeable. Many of my studies could not be continued and accomplished without his insightful input and advice. From him, I saw the diligence and the persistence a devoted scientist should have. I will never forget how patient he was with me and how he assisted me with scientific writing. His commitment in training me as a good scientist will definitely be extremely helpful for my future academic career.

My thanks also go to my wonderful labmates. My works benefit greatly from the brainstorming with Xionglei He, Peng Shi, Margaret Bakewell, Ondrej Podlaha and Wenfeng Qian. Wendy Grus and Margaret Bakewell were my English teachers and the editors for many of my poorly written drafts. Soochin Cho is the one who always listened and understood. Zhi Wang and Zhihua Zhang assisted me in many computational projects, and provided jokes almost hourly. These friends made my doctoral study an enjoyable and unforgettable journey.

I also wish to thank Patricia Wittkopp, Zhaohui Steve Qin and Priscilla Tucker, the other three members of my dissertation committee, for discussing my data, commenting on earlier drafts of the dissertation and my other studies, and assisting me in fellowship and job applications.

Finally, I would like thank my family, especially my wife Yu-Ling Lu, for the unconditional love and support through out this entire venture. Yu-Ling's devotion to the family and companionship in this foreign country give me an enviable life as a graduate student.

This research is partially funded by EEB department fellowship and Rackham Predoctoral Fellowship from University of Michigan.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Comparing the expression-profiles of over 10,000 genes from the human and mouse genomes, I address fundamental questions on mammalian gene expression. First, I demonstrate that over 80% of human-mouse orthologous genes are evolutionarily conserved in their expression-profiles. This result highlights the importance of proper gene expression to fitness. Second, I show that highly expressed and tissue-specific genes tend to evolve slowly in expression-profile, implying that the expression pattern is of particular importance to highly expressed and tissue-specific genes. I then investigate the potential roles that gene expression plays in protein sequence evolution, dynamics of genome organization, and evolutionary changes of gene essentiality in mammals. My results indicate that tissue-specificity is a stronger determinant on protein evolutionary rate than gene expression level, a factor that is known to be the most important rate determinant in yeasts. The result suggests a great variation in rate determinants of protein sequence evolution between unicellular and multicellular organisms. Subsequently, my analyses on the origin of co-expressed gene clusters indicate that co-expression of linked genes is a form of transcriptional interference that is disadvantageous to organisms, suggesting that transcriptional interference may promote recurrent relocations of genes in the genome. Lastly, I study underlying mechanisms of the evolution of gene essentiality. The results show that the changes of gene essentiality appear to be associated with adaptive evolution at the protein-sequence level, while gene duplication and gene expression evolution plays a negligible role. Together, my studies help understand patterns, mechanisms and consequences of gene expression evolution.

INTRODUCTION

Understanding the molecular basis of organismal evolution is one of the major tasks of biology. For example, many scientists have been trying to identify the genetic bases underlying the major transitions of life forms throughout the history and to understand the causes setting us apart from other primates (Olson 1999; Zhang, Webb, and Podlaha 2002; Zhang 2003; Hayakawa et al. 2005; Prabhakar et al. 2006; Calarco et al. 2007; Harris, Rogers, and Milosavljevic 2007). In evolution, gene function can be altered through changes in either protein function or gene expression. Several case studies clearly demonstrated that gene expression changes can result in phenotypic changes with significant evolutionary ramifications. For instance, the variation in beak depth and breadth among Darwin's finches (*Geospiza sp.*) is due to the variation in *Bmp4* gene expression pattern (Abzhanov et al. 2004). In humans, *cis*-regulatory changes elevating lactase (*LCT*) transcription enables northern Europeans to digest lactose through adulthood (Bersaglieri et al. 2004), while >90% of Asian people experience the condition of lactose intolerance.

In spite of the important role gene expression evolution may play in organismal evolution, previous research mainly focused on protein coding-sequence evolution (Li 1997; Nei and Kumar 2000). Because of that, several fundamental questions associated with gene expression evolution, such as how gene expression changes during evolution, how regulation of gene expression originated, and how expression evolution connects genomic and phenotypic evolution, are largely unexplored. Studies on gene evolution at the

transcriptional level can provide a cohesive view of organismal evolution from DNA to phenotypes.

Until recently, the technical difficulty associated with simultaneously measuring the expression of a large number of genes was a bottleneck for studies of general patterns of gene expression. Nowadays, high-throughput microarray technologies allow the measurement of gene expression at the genomic scale, resulting in the generation of numerous genome-wide gene expression data for many organisms under various conditions. These data enable me to use computational approaches to address several fundamental questions that have puzzled biologists for years. My dissertation addresses gene expression evolution in mammals. I approach these questions by comparing the expression-profiles (i.e., relative gene expression levels across different cell types) of over 10,000 genes from the human and mouse genomes (Su et al. 2004).

My dissertation is comprised of two main sections. The first section is on the evolutionary constraints and patterns of mammalian gene expression. In Chapter 1, I reject the completely neutral model of transcriptome evolution (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004) by confirming the evolutionary conservation of expression-profile between >80% of human-mouse orthologs. In Chapter 2, I study the evolutionary rate of gene expression-profiles. The result suggests that expression-profile is of particular importance and tends to be more evolutionarily conserved for highly-expressed gene and tissue-specific genes.

In the second section, I describe the potential roles gene expression has in molecular and organismal evolution at different levels. The subjects in Chapter 3, Chapter 4 and

Chapter 5 are the effects of gene expression on protein sequence evolution, dynamics of genome architecture, and gene knockout phenotype evolution, respectively.

In Chapter 3, I compare the relative importance of the rate determinants for mammalian protein evolution, including two important properties of gene expression, namely expression level and tissue-specificity. I found considerable differences in the rules governing protein evolution between yeast and mammals. In Chapter 4, I focus on the origin of co-expressed gene clusters in eukaryotic genomes. Previous authors assumed that this phenomenon is a result of adaptive relocation of initially unlinked but co-expressed genes (e.g. Hurst, Pal, and Lercher 2004). I propose and test an alternative hypothesis that co-expression of linked genes is a form of transcriptional interference that is disadvantageous to the organism. My result suggests that transcriptional interference may underlie recurrent relocations of genes in the genome. In Chapter 5, I investigate the molecular mechanisms responsible for the observation that 20% of mouse orthologs of human essential genes are non-essential. Here, an essential gene is defined by its knockout phenotype of premature death or sterility. Three possible mechanisms are examined: (i) functional compensation between duplicate genes, (ii) divergence of protein sequences, and (iii) divergence of gene expression. I found that the changes of gene essentiality appear to be associated with adaptive evolution at the protein-sequence level.

LITERATURE CITED

Abzhanov A, Protas M, Grant BR, Grant PR, and Tabin CJ. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science **305**:1462-1465.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, and Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet **74**:1111-1120.

Calarco JA, Xing Y, Caceres M, Calarco JP, Xiao X, Pan Q, Lee C, Preuss TM, and Blencowe BJ. 2007. Global analysis of alternative splicing differences between humans and chimpanzees. Genes Dev **21**:2963-2975.

Harris RA, Rogers J, and Milosavljevic A. 2007. Human-specific changes of genome structure detected by genomic triangulation. Science **316**:235-237.

Hayakawa T, Angata T, Lewis AL, Mikkelsen TS, Varki NM, and Varki A. 2005. A human-specific gene in microglia. Science **309**:1693.

Hurst LD, Pal C, and Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet **5**:299-310.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, and Paabo S. 2004. A neutral model of transcriptome evolution. PLoS Biol **2**:682-689.

Li W-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Nei M, and Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet **64**:18-23.

Prabhakar S, Noonan JP, Paabo S, and Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science **314**:786.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, and Hogenesch JB. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Yanai I, Graur D, and Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15-24.

Zhang J. 2003. Evolution of the human ASPM gene, a major determinant of brain size. Genetics **165**:2063-2070.

Zhang J, Webb DM, and Podlaha O. 2002. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. Genetics **162**:1825-1835.

CHAPTER 1

EVOLUTIONARY CONSERVATION OF MAMMALIAN GENE EXPRESSION: THE STUDY OF HUMAN-MOUSE ORTHOLOGOUS GENES

1.1 ABSTRACT

Mouse models are often used to study human genes, because it is believed that the expression and function are similar for the majority of orthologous genes between the two species. However, recent comparisons of microarray data from thousands of orthologous human and mouse genes suggested rapid evolution of gene expression profiles under minimal or no selective constraint. These findings appear to contradict non-array-based observations from many individual genes and imply the uselessness of mouse models for studying human genes. Because absolute levels of gene expression are not comparable between species when the data are generated by species-specific microarrays, use of relative mRNA abundance among tissues (*RA*) is preferred to that of absolute expression signals. We thus reanalyze human and mouse genome-wide gene expression data generated by oligonucleotide microarrays. We show that the mean correlation coefficient among expression profiles detected by different probe sets of the same gene is only 0.38 for human and 0.28 for mouse, indicating that current measures of expression divergence are flawed because the large estimation error (discrepancy in expression signal detected by different probe sets of the same gene) is mistakenly included in the between-species divergence. When this error is subtracted, 84% of human-mouse orthologous gene pairs show significantly lower expression divergence than that of random gene pairs. In contrast to a previous finding, but consistent with the common sense, expression profiles of orthologous tissues between species are more

similar to each other than to those of non-orthologous tissues. Furthermore, the evolutionary rate of expression divergence and that of coding sequence divergence are found to be weakly, but significantly positively correlated, when *RA* and the Euclidean distance are used to measure expression-profile divergence. These results highlight the importance of proper consideration of various estimation errors in comparing the microarray data between species.

1.2 INTRODUCTION

Patterns and mechanisms of DNA and protein sequence evolution have been extensively studied in the past three decades (Li 1997; Nei and Kumar 2000). By contrast, little had been known about the general patterns of gene expression evolution until a few years ago when high-throughput gene expression profiling technologies became available (Cavalieri, Townsend, and Hartl 2000; Enard et al. 2002; Oleksiak, Churchill, and Crawford 2002; Ranz et al. 2003; Rifkin, Kim, and White 2003; Townsend, Cavalieri, and Hartl 2003). Because gene expression evolution links the evolutionary changes of genes and phenotypes, it is of fundamental importance to estimate the rate of gene expression evolution and to look for the underlying molecular mechanisms responsible for transcriptome evolution. Among all the technologies for producing transcriptome data, the DNA (oligonucleotide or cDNA) microarray is most commonly used (Ruan et al. 2004). This technology makes it possible to study the expression of large numbers of genes simultaneously with relatively low cost compared to sequencing-based technologies such as serial analysis of gene expression (SAGE) (Velculescu et al. 1995) and massively parallel signature sequencing (MPSS) (Brenner et al. 2000). Transcriptome data from various tissues in various organisms have

been produced using DNA microarrays, making evolutionary analysis of genome-wide gene expression patterns feasible.

Based on oligonucleotide microarray datasets obtained from the human and mouse (Su et al. 2002), Yanai, Graur, and Ophir (2004) found that the expression profiles of orthologous genes differ substantively between the two species, suggesting little selective constraint in the evolution of gene expression. Additionally, based on the expression similarity among 32 human and mouse tissues, they found orthologous tissues between species (e.g., human liver and mouse liver) to be less similar than non-orthologous tissues within species (e.g., human liver and human testis). Because tissue functions are determined by the genes expressed in the tissue, these results imply that the human liver is functionally more similar to the human testis than to the mouse liver, which is contrary to the common sense. Based on both oligonucleotide and cDNA microarrays, Khaitovich et al. found that expression-level divergence between primate species increases linearly with divergence time and that functional genes and expressed pseudogenes have similar rates of expression evolution (Khaitovich et al. 2004). Because pseudogenes evolve without any selective constraint, these results suggest that gene expression evolution is largely neutral, without the influence of either positive or purifying selection (Khaitovich et al. 2004).

These findings are surprising for several reasons. First, it is well established that coding sequences and functions of most orthologous genes are conserved across species (Li 1997). Because a gene must be expressed properly to function in the cell, it is puzzling why gene function should be conserved when the expression changes quickly. Second, the expression pattern and function of human genes are often inferred from their mouse orthologs (e.g., Hinds et al. 1993), based on the assumption that these properties are

conserved between the two species. The success of many mouse models of human genes and diseases suggests the validity of this assumption. Studies using traditional non-array-based methods such as the Northern analysis showed that the expression profiles of human-mouse orthologs are overall similar, although a quantitative genome-wide measure of mean similarity is difficult to obtain, due to large variations in experimental designs among these individual-gene studies. Third, a recent microarray-based study of the nematode *Caenorhabditis elegans* showed that transcriptome evolution is significantly faster in lab mutation-accumulation strains than in naturally isolated strains (Denver et al. 2005). Because the sizes of the lab populations are much smaller than those of the wild populations, Denver et al's observation is best explained by purifying selection acting on expression divergence in nature. The rate of expression divergence would have been similar between lab and wild populations if gene expression were not under any selection (Kimura 1983).

With these considerations, we reexamined the expression divergence between orthologous genes, based on the oligonucleotide microarray data of human and mouse genes generated by Su et al. (2004). Our choice of this dataset is not only because it (or its earlier version) has been used in a number of evolutionary studies (Li 1997; Makova and Li 2003; Huminiecki and Wolfe 2004; Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005; Yang, Su, and Li 2005), but also because this dataset is one of the largest for humans and mice, the expression divergence between which is of special importance to the biomedical community. Our analysis showed that there is a large error in measuring gene expression using microarrays. When this error is subtracted, the majority of orthologous genes show significantly lower expression-profile divergence between humans and mice than expected under neutrality.

1.3 MATERIALS AND METHODS

1.3.1 Mapping expression data to Ensembl genes

We used the Gene Atlas V2 microarray dataset of humans and mice (http://symatlas.gnf.org/). The dataset was generated by hybridization of RNA from 79 human and 61 mouse tissues onto the designed Affymetrix microarray chips (human: U133A/GNF1H; mouse: GNF1M) (Su et al. 2004). These human and mouse chips were designed according to the annotated human and mouse genome sequences, respectively. On a chip, each gene is represented by one to several probe sets, each of which comprises 11 pairs of probes. Each pair of probes contains a oligonucleotide probe that matches the genomic sequence perfectly and the second probe that is identical to the first probe except for the middle nucleotide, which differs from the genomic sequence. To assign the probe sets to their corresponding human or mouse genes, we aligned sequences of each probe set to the Ensembl cDNA sequences (human: Homo_sapiens.NCBI35.feb.cdna.fa; mouse: Mus_musculus.NCBIM33.feb.cdna.fa; http://www.ensembl.org/) using BLASTn (http://www.ncbi.nlm.nih.gov/blast/) and retained those probe sets in which all matching probes perfectly matched to the same Ensembl gene. 25,368 probe sets (75.3%) in the human chip corresponding to 16,456 genes and 18,005 probe sets (49.8%) in the mouse chip corresponding to 15,835 genes were used for further analysis. The expression level detected by each probe set was obtained as the signal intensity (S) computed by the MAS 5.0 algorithm (Hubbell, Liu, and Mei 2002). The S values were averaged among replicates.

To compare our results with those of Yanai, Graur, and Ophir (2004), we also analyzed the expression dataset used in their study, which is from Su et al. (2002). The results obtained from the two datasets were consistent with each other when the same method

was used. We also used the expressional values computed by the robust multi-array averaging (RMA) (Bolstad et al. 2003; Irizarry et al. 2003). The results obtained from MAS 5.0 and RMA algorithms were similar. Thus, only the results derived from the dataset Gene Atlas V2 and calculated by MAS 5.0 are presented here.

1.3.2 Human-mouse orthologs

Homology information of human and mouse genes was obtained from Ensembl EnsMart (http://www.ensembl.org/Multi/martview) (Kasprzyk et al. 2004). There are several homologous relationships between human and mouse genes annotated by Ensembl. We only considered those pairs of genes annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BLASTZ searches) as orthologous. 10,607 pairs of human-mouse orthologs have expression information from the microarray data we use. Among these genes, 64.5% of human genes and 86.9% of mouse genes were represented by a single probe set, while the others have multiple probe sets on the chips. Affymetrix probes with name suffixes "_x_at" and "_s_at" are believed to be prone to cross-hybridization, compared to other probes (Affymetrix Technical Support, Data Analysis Fundamentals, Appendix B; http://www.affymetrix.com/), and have been considered "suboptimal" (Huminiecki and Wolfe 2004; Yang, Su, and Li 2005). But our analysis showed that the quality of these probes is not worse than other probes (see below). We therefore considered all probe sets equally.

The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse

orthologs were retrieved form Ensembl EnsMart. In this database, d_S and d_N were estimated by codeml of the PAML package (Yang 1997) using the likelihood method.

1.3.3 Analysis of gene expression data

For the purpose of studying the divergence of expression profiles between human-mouse orthologs, we analyzed 26 common tissues from the two species (see Figure 1.2 for the tissues examined; note that mouse lower spinal cord was used as the homologous tissue of human spinal cord).

For calculating the expressional divergence between a pair of orthologous genes, the expression signal in human and that in mouse must be comparable. Log_2-transformed signal intensity (*S*) is commonly used to quantitatively measure the level of gene expression. But it has intrinsic problems for comparing expression data derived from different Affymetrix microarrays. First, probes are separately designed for the human and mouse orthologous genes and do not target the same sequences. Therefore, the human probes and mouse probes have different affinities to their target RNAs (Binder et al. 2004a; Binder et al. 2004b). Subsequent normalization procedures still depend on the properties of the microarray chip and do not make the expression signals of orthologous genes from different chips comparable. This important technical issue was apparently ignored in earlier evolutionary studies (Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005), as the authors directly compared *S* or log_2-transformed *S* values obtained from the human and mouse chips. Second, because the *S* value detected by the microarray is approximately linear with the actual quantity of target RNA within reasonable ranges of measurement (Affymetrix 2001), log_2-transformed *S* values tend to overestimate the difference between two low expressional

values but underestimate the difference between two high expressional values. For these two reasons, we used relative abundance (*RA*) to measure the relative expression level of a gene in a given tissue among the sampled tissues. The *RA* for human or mouse gene *i* expressed in tissue *j* is defined as

$$RA_H(i,j) = S_H(i,j) / \sum_{j=1}^{n} S_H(i,j) \text{ and } RA_M(i,j) = S_M(i,j) / \sum_{j=1}^{n} S_M(i,j). \quad (1)$$

Here, *n* is the number of common tissues considered and is 26 in this study. *H* indicates human and *M* indicates mouse. $S_H(i,j)$ and $S_M(i,j)$ are the expression signal intensities of gene *i* in human tissue *j* and mouse tissue *j*, respectively. When the RMA algorithm was used to measure expression, *S* was calculated by anti-log of the default output value. It should be noted that by using *RA* we lose the information of the absolute expression level in all tissues, but as aforementioned, the absolute expression levels of orthologs are practically incomparable.

The divergence between expression profiles of human and mouse orthologous genes is measured by "1 - Pearson's correlation coefficient (*r*)" and "Euclidean distance (*d*)" using the *RA* values of the 26 pairs of tissues. Pearson's *r* between human and mouse gene *i* is computed by:

$$r = \frac{\sum_{j=1}^{n}[RA_H(i,j)RA_M(i,j)] - [\sum_{j=1}^{n} RA_H(i,j)][\sum_{j=1}^{n} RA_M(i,j)]/n}{\sqrt{\sum_{j=1}^{n}[RA_H(i,j)]^2 - [\sum_{j=1}^{n} RA_H(i,j)]^2/n}\sqrt{\sum_{j=1}^{n}[RA_M(i,j)]^2 - [\sum_{j=1}^{n} RA_M(i,j)]^2/n}}. \quad (2)$$

Euclidean distance *d* between human and mouse gene *i* is computed by:

$$d = \sqrt{\sum_{j=1}^{n}[RA_H(i,j) - RA_M(i,j)]^2}. \quad (3)$$

For those genes with more than one probe set, we randomly pick one probe set to represent that gene while measuring the expressional divergence of human-mouse orthologs.

Our analysis showed that different probe sets on the same chip often give very different *S* values for a given gene in a given tissue. This difference is most likely due to the variation in affinity among probe sets for a given gene. Let d_H denote the Euclidean distance between the expression profiles estimated by two randomly picked probe sets for the same human gene, d_M denote the Euclidean distance for the corresponding mouse gene, and d be the Euclidean distance defined in formula [3]. We estimate the net distance (D) between human and mouse orthologous genes by

$$D = d - (d_H + d_M)/2. \quad (4)$$

This procedure is analogous to the estimation of the net genetic difference between two populations by subtracting the genetic variation within populations from that between populations (Nei 1987). D should be interpreted as the detectable divergence in expression-profile between an orthologous gene pair from the human and mouse. A lower d than $(d_H + d_M)/2$ indicates no detectable divergence under the current technology; we therefore assume D=0. There were 3762 and 1385 genes with multiple probe sets on the human and mouse chip, respectively. The number of genes with multiple probe sets in at least one species is 4564. These 4564 genes were used to measure D. For a given gene, when one of the two species has only one probe set, we assumed $d_H = d_M$. Human and mouse genes in this set of 4564 orthologs were randomly paired to estimate the neutral expectation of expression divergence between the two species.

The tissue expression dendrograms were calculated from the matrix of distances among tissues, which were estimated from the *RA* values of 10,607 gene pairs in humans and

mice. Pearson's correlation coefficient between human tissue *j*1 and mouse tissue *j*2 is computed by

$$r(human\ j1_1, mouse\ j2) = \frac{\sum_{i=1}^{N}[RA_H(i, j1)RA_M(i, j2)] - [\sum_{i=1}^{N} RA_H(i, j1)][\sum_{i=1}^{N} RA_M(i, j2)] / N}{\sqrt{\sum_{i=1}^{N}[RA_H(i, j1)]^2 - [\sum_{i=1}^{N} RA_H(i, j1)]^2 / N}\sqrt{\sum_{i=1}^{N}[RA_M(i, j2)]^2 - [\sum_{i=1}^{N} RA_M(i, j2)]^2 / N}} \quad (5)$$

Here *N* is the total number of genes studied. The correlation coefficient between tissue *j*1 and *j*2 of the same species, say human, can be computed by replacing *M* with *H* in the above formula. Distance between two tissues is defined as 1-*r*. The dendrograms of tissues were derived from the hierarchical clustering algorithm (Murtagh 1985) implemented in R (http://www.r-project.org/). We generated 10,000 dendrograms by bootstrapping genes. For a gene with more than one probe set, we randomly pick a probe set whenever the gene is sampled. The final consensus tree was constructed by MEGA3 (Kumar, Tamura, and Nei 2004). In addition to the distance 1-*r*, we also used the Euclidean distance to compute the distance of expression profiles between two tissues. That is, the Euclidean distance between human tissue *j*1 and mouse tissue *j*2 is computed by

$$d(human\ j1, mouse\ j2) = \sqrt{\sum_{i=1}^{N}[RA_H(i, j1) - RA_M(i, j2)]^2} \quad (6)$$

The distance between tissue *j*1 and *j*2 of the same species, say human, can be computed by replacing *M* with *H* in the above formula.

1.4 RESULTS AND DISCUSSIONS

1.4.1 The mean expression divergence between human-mouse orthologs is lower than that between random gene pairs

To study whether the expression profile is conserved between human and mouse orthologous genes, it is necessary to know the expected value of expression divergence under complete neutrality. Ideally, this value should be estimated using expressed pseudogenes. However, it is unlikely that a functionless pseudogene generated before the separation of primates and rodents would still be retained and expressed in humans and mice. Jordan, Marino-Ramirez, and Koonin (2005) suggested that the expected expression divergence under neutrality can be approximated by the expression difference between a randomly picked human gene and a randomly picked mouse gene. This suggestion was based on the assumption that the expression similarity between human-mouse orthologs has been completely lost under neutrality. Following this logic, we compared expression divergence of orthologous human and mouse genes with that of randomly paired human and mouse genes. Our dataset included 10,607 orthologous gene pairs, each with expression information from 26 common tissues from the two species (see Materials and Methods). For a gene of a species, the signal intensity (*S*) values from all the tissues were transformed to the relative abundance (*RA*) values, which are the signal intensities in individual tissues divided by the total signals in all the tissues considered (see formula [1] in Materials and Methods). The reason for using *RA*, instead of *S*, is that the absolute expression signal is not meaningful in the human-mouse comparison as different microarrays were used for the two species. We then used 1 - Pearson's correlation coefficient (*r*) to measure the divergence of expression profile between species (see formula [2] in Materials and Methods). Note that *r* can vary between -1 and 1, thus the value of 1-*r* ranges from 0 to 2. Higher *r* means lower 1-*r* and indicates smaller expression divergence. For randomly paired human-mouse genes, the mean *r* is 0.002 and the median *r* is -0.024 (Figure 1.1a). Hence, on average, randomly paired

human and mouse genes show no expression similarity, as expected. For human and mouse orthologous gene pairs, the mean *r* is 0.216 and the median *r* is 0.165. Orthologous genes have significantly higher *r* values than random gene pairs do ($P<10^{-280}$, Mann-Whitney U test). Thus, the expression divergence between human and mouse orthologous genes is on average lower than expected under complete neutrality, suggesting that expression evolution has been under purifying selection. We also used another commonly used measure of expression divergence, Euclidean distance (*d*) (see formula [3] in Materials and Methods), which is always positive. Higher *d* indicates greater expression divergence. We found that human-mouse orthologs have a mean *d* of 0.180 and a median *d* of 0.149 (Figure 1.1b). By contrast, human-mouse random gene pairs have a mean *d* of 0.229 and a median *d* of 0.177 (Figure 1.1b). Again, expression divergence measured by Euclidean distances is lower between orthologous gene pairs than between random pairs ($P<10^{-151}$, Mann-Whitney U test), consistent with the result obtained from the use of *r*. These results support the findings of Yanai, Graur, and Ophir (2004) and Jordan, Marino-Ramirez, and Koonin (2005), although they used different datasets and/or different ways of analysis. For example, Yanai, Graur, and Ophir (2004) used an earlier version of the microarray data (Su et al. 2002), which was considerably smaller than the version used here. They also removed those genes with more than one probe set on the chip. The number of orthologous gene pairs they analyzed was only 1350, compared with 10,607 in our analysis. When multiple probe sets are available for a gene, Jordan, Marino-Ramirez, and Koonin (2005) picked the probe set that showed highest *S*, while we randomly picked a probe set. Because it is unclear whether the use of the probe set with highest expression signal introduces biases, our result should be less influenced by such potential biases.

1.4.2 Errors in measuring gene expression

Although the overall rate of transcriptome evolution between humans and mice appears lower than the neutral expectation, it is unclear what proportion of genes are under purifying selection in their expression evolution and how strong the selection is. To address this question, it is necessary to evaluate the error in measuring gene expression by microarrays. It is quite common that more than one probe set is used to represent a gene on an oligonucleotide microarray. Theoretically, if the same transcripts are targeted and if there is no cross-hybridization, all probe sets designed for the same gene should provide the same, or at least similar, expression signals. However, this is often not the case. For example, there are two probe sets on the human chip and two on the mouse chip for the gene *RUTBC1*. The *S* as well as *RA* values obtained from the two probe sets on the same chip are quite different even for the same tissues (Figure 1.2). In this example, the *r* between the *RA* values generated from the two mouse probe sets (0.23) is even lower than the average *r* between the *RA* values generated from a human probe set and a mouse probe set (0.40). In other words, the apparent low *r* between species is largely attributable to the estimation error of gene expression within species. For many of the 3762 genes with multiple probe sets on the human chip, the *r* values between the expression profiles generated by two randomly picked probe sets of the same gene are much lower than 1 (Figure 1.3a). In fact, *r* has a mean of 0.375 and a median of 0.368. There are 1385 genes with multiple probe sets on the mouse chip. The mean *r* is 0.277 and the median *r* is 0.235 between the expression profiles generated by two randomly picked probe sets of the same mouse gene (Figure 1.3a). These low *r* values show that the expression level is not precisely measured by the microarrays. Rather, there are large errors associated with the estimates. Similar results were obtained

when *d* was used to measure the difference between expression profiles detected by different probe sets targeting the same gene (Figure 1.3b).

It is unclear what factors caused such a great variation between expressional levels detected by different probe sets. Affymetrix probes with name suffixes "_x_at" and "_s_at" are thought to be prone to cross-hybridization, compared to other probes, and have been considered "suboptimal" (Huminiecki and Wolfe 2004; Yang, Su, and Li 2005). Using the human chip, we examined whether the high estimation error is due to the inclusion of suboptimal probe sets. Let A be the group of genes with multiple probe sets and B be the subset of A that contain multiple optimal probe sets but no suboptimal probe sets. There are 3762 genes in group A and 1097 genes in group B. Assuming that "suboptimal" probes are randomly distributed among genes, we expect that the mean expression-profile similarity between two probe sets of the same gene is higher for group B genes than for group A genes. However, we observed the opposite pattern (Figure A.1), suggesting that the estimation error was not due to the inclusion of "suboptimal" probe sets. We also compared the average *S* of all 26 tissues generated by "suboptimal" probe sets and the corresponding value generated by "optimal" probe sets for each of the human genes with at least one "suboptimal" and at least one "optimal" probe sets. We found that 1,016 genes show higher mean *S* from "suboptimal" probe sets than from "optimal" probe sets and that 1,187 genes exhibit the opposite pattern ($P<0.0003$, chi-squares test). Thus, although "suboptimal" probe sets tend to show slightly lower signal intensities than "optimal" probe sets do, the bias is small. Furthermore, in the presence of cross-hybridization, it is unclear whether our observation means that "suboptimal" probes are less or more accurate than "optimal" probes. In sum, we found no definite evidence that "suboptimal" probe sets performed worse than "optimal"

probe sets. Other possible sources of the estimation error are cross-hybridization with products of multiple genes, hybridization with different transcripts of the same gene, variable hybridization affinities of different probe sets, and stochastic background noise of the microarray. To address estimation errors, a variety of methods have been employed, such as removing "suboptimal" probe sets (Huminiecki and Wolfe 2004; Yang, Su, and Li 2005), discarding genes with multiple probe sets (Yanai, Graur, and Ophir 2004), and selecting the probe set with the highest expression level for each gene (Jordan et al. 2004; Jordan, Marino-Ramirez, and Koonin 2005). However, none of these strategies remove the estimation error. First, as we have shown, eliminating "suboptimal" probe sets does not reduce the estimation error. Second, removing genes with multiple probe sets does not reduce the intrinsic imprecision of individual probe sets. Finally, choosing the highest signal may only alleviate the error slightly because it is still unknown whether the quality of the highest-signal probe set on the human chip is comparable to that on the mouse chip and whether highest-signal means most accurate.

1.4.3 Expression profiles of 84% of human-mouse orthologs are significantly lower than expected under neutrality

Despite the high estimation error shown in Figure 1.3, the expression differences between human and mouse orthologs are higher than those detected by different probe sets within species. This indicates that for many genes the expression profile is not completely conserved between the two species. For estimating the proportion of human-mouse orthologs that diverge significantly slower than expected under neutrality, Euclidean distance d is preferred over $1-r$, because the correlation coefficient r ignores any linear changes which may exist between expression profiles. We computed the net expression distance D between

human and mouse by subtracting the expression distance between probe sets within species from the expression distance between species (see formula [4] in Materials and Methods). This procedure is analogous to the estimation of the net genetic difference between populations by subtracting the variation within populations (Nei 1987). *D* can be interpreted as the detectable expression divergence given the estimation error. Randomly paired human-mouse genes should have no expression similarity, thus the Euclidean distances do not require correction. We found that *d* has a relatively wide distribution for random gene pairs (Figure 1.4). Five percent of *d* are smaller than $d_{5\%}$=0.0897. If the *D* value of a human-mouse orthologous gene pair is smaller than $d_{5\%}$, we may claim that the expression divergence of this gene pair has been under selective constraint because the probability that the evolution has been neutral is lower than 5% (Figure 1.4). Using this criterion, we found that the detectable expression divergence of 83.9% of genes is significantly lower than expected under complete neutrality. A simple interpretation is that the expressions of these genes are under purifying selection. However, our result may also reflect the large estimation error of the current microarray technology and consequently low detectable expression divergence between species. More accurately, our findings suggest that at least for 84% of genes the current data do not provide evidence for neutrality. Note that this estimate was derived from 4564 orthologous gene pairs in which multiple probe sets are available for at least one species so that the estimation error could be evaluated. Under the assumption that the probe design for a gene is independent of the rate of gene expression evolution, our result is applicable to the entire genome. We also computed the values of d_H and d_M by averaging the Euclidean distances of all possible combinations of probe sets of the

same gene, instead of using two randomly picked probe sets. The results are very similar (Figure A.2).

1.4.4 Orthologous tissues between species are more similar than non-orthologous tissues in terms of expression profile

It is expected that orthologous tissues between species (e.g., human liver and mouse liver) should have similar expression profiles because they carry out similar physiological functions. However, Yanai, Graur, and Ophir (2004) showed that 16 pairs of human and mouse tissues were clustered into two species-specific clades. They explained that the dichotomy of human and mouse tissue expression in the dendrogram is due to large numbers of changes in the expression programs and considered this pattern as evidence for the neutral evolution of transcriptional profiles. To understand the exact cause of their findings, we regenerated the tissue expression dendrograms. When we measured the gene expression level by the normalized RMA default output or $\log_2$ MAS 5.0 signal intensity as Yanai, Graur, and Ophir (2004) did, we reproduced their dendrogram that showed the separate clustering of human and mouse tissues. However, a completely different dendrogram is produced when *RA* is used. Now most of the human-mouse orthologous tissues cluster, with bootstrap values higher than 95% (Figure 1.5). Similar results were obtained when either 1-*r* (Figure 1.5a) or *d* (Figure 1.5b) was used to measure the expression distance between tissues (formulas [5] and [6] in Materials and Methods) or when a smaller microarray dataset (Su et al. 2002) was used. Because our results are more consistent with the expectation, we believe that the previous observation by Yanai, Graur, and Ophir (2004) was due to inappropriate data processing. In particular, ignoring the systematic bias caused by the use of two different

oligonucleotide arrays and direct comparison of *S* (or log-transformed *S*) between species was the major problem (see Materials and Methods). Irrespective of the distance measure used, some non-orthologous species-specific tissue clusters remain in our dendrograms with relatively high bootstrap percentages (Figure 1.5). For example, amygdala, hypothalamus, and spinal cord of the same species cluster together. Whether this phenomenon is owing to simultaneously rapid evolution of the genes co-expressed in these regions or other reasons requires further investigation.

1.4.5 Correlation between the rate of expression-profile divergence and that of coding sequence divergence

It has been controversial as to whether there is a positive correlation between the rate of expression evolution and the rate of coding sequence evolution across many genes in a genome. In earlier studies, this question was addressed by comparing duplicate genes within species (Wagner 2000; Gu et al. 2002; Makova and Li 2003). However, such analysis can only test whether the expression divergence and sequence divergence between two homologous genes are correlated, but cannot test whether the rates of the two divergences are correlated. The latter question can be answered only when orthologous genes between two species are compared. Recent studies using human-mouse orthologs, however, do not find such a correlation (Jordan et al. 2004; Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005). In contrast, we found significantly positive correlations between *d* and coding sequence divergence in terms of d_S (Pearson's correlation coefficient=0.119, $P<10^{-32}$; Fig. 6a) or d_N (Pearson's correlation coefficient =0.187, $P<10^{-80}$; Fig. 6b). Because all the human-mouse orthologous gene pairs have the same divergence

time, these results show that the rate of gene expression-profile divergence and the rate of coding sequence divergence at both the synonymous and nonsynonymous sites are positively correlated. In other words, proteins with low rates of sequence evolution also tend to have low rates of expression-profile evolution. We also observed that d and d_N/d_S are positively correlated (Pearson's correlation coefficient=0.152, $P<10^{-52}$; Fig. 6c). In our dataset, there is virtually no correlation between d_N/d_S and d_S (Pearson's correlation coefficient=0.023, P=0.02), contrary to the observation of a high positive correlation in a recent analysis of 3561 human-mouse orthologous gene pairs (Wyckoff et al. 2005). Thus, d_N/d_S is a reliable measure of the strength of purifying selection acting on coding sequences. The correlation between d and d_N/d_S suggests that the level of purifying selection preventing protein sequence divergence is positively correlated with the level of purifying selection preventing expression-profile divergence. We think that this correlation arises because when a gene is functionally important, both its protein sequence and expression profile tend to be more conserved, compared to the situation when the gene is less important. It should be pointed out that although the correlations we detected are highly significant, the magnitudes of the correlations are low. This is not unexpected, as both the rate of sequence evolution and that of expression-profile evolution are determined by multiple factors and the estimated rate of expression-profile evolution has large errors. A recent study of humans and chimpanzees suggested that the rate of gene sequence evolution and the rate of expression-level evolution may also be positively correlated (Khaitovich et al. 2005).

We believe that the main reason why the positive correlation between expression-profile divergence and coding sequence divergence was not previously observed in the comparison of human and mouse orthologous genes (Yanai, Graur, and Ophir 2004; Jordan,

Marino-Ramirez, and Koonin 2005) is because *S* (or the log-transformed *S*) was used, instead of *RA*, in the estimation of *d*. However, the observed difference in the absolute level of gene expression between species is not meaningful due to the use of species-specific microarrays. Therefore, inclusion of this difference in computing *d* substantially increases the noise. Furthermore, highly expressed genes tend to be more conserved at the coding sequence level (Pal, Papp, and Hurst 2001; Rocha and Danchin 2004; Jordan, Marino-Ramirez, and Koonin 2005; Zhang and He 2005). They also tend to have high *d* when it is computed using *S* (Jordan, Marino-Ramirez, and Koonin 2005). Together, these factors dramatically reduce the positive correlation between *d* and coding sequence divergence. This problem is rectified when *d* is computed by *RA*.

As mentioned, 1-*r* and *d* are commonly used to measure gene expression divergence. Compared to *d*, 1-*r* is more often adopted by evolutionists, such as in the studies of duplicated genes in yeast (Wagner 2000; Gu et al. 2002), nematode (Castillo-Davis, Hartl, and Achaz 2004; Conant and Wagner 2004), human (Makova and Li 2003; Huminiecki and Wolfe 2004), mouse (Huminiecki and Wolfe 2004), and mustard (Haberer et al. 2004). However, *d* reportedly performs better (Slonim et al. 2000) and has been used to compare orthologous genes (Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005) and to cluster co-expressed genes (Wen et al. 1998; de Bivort, Huang, and Bar-Yam 2004). To our surprise, we did not observe positive correlations between 1-*r* and either d_N, d_S, or d_N/d_S. But this result is consistent with that of Jordan et al. (2004), although they used Spearman's rank correlation instead of Pearson's correlation (*r*) to measure the human-mouse expression-profile similarity. One may think that the expression divergences measured by *d* and 1-*r* should be positively correlated. However, the mathematical

properties of *d* and 1-*r* are different. For example, any linear transformations of *S* do not affect *r*, while they may influence *d*. In addition, 1-*r* is bounded between 0 and 2, whereas *d* can increase infinitely. In our data, 1-*r* and *d* have a weak, but significant, negative correlation (Figure A.3). Although both measures are commonly used, which one better describes the expression divergence between orthologous genes remains unanswered. It is possible that the advantages of these two measures vary depending on the conditions used. It is also important to note that neither 1-*r* nor *d* measures the number of genetic changes (i.e., number of nucleotide substitutions) underlying the observed expression-profile divergence. Because the molecular mechanism of gene expression regulation is complex and not well understood, no distance measures currently exist to quantify the genetic changes underlying expression-profile divergence (Leung and Cavalieri 2003).

1.4.6 Final remarks

There are two ways to compare the transcriptomes of two species using DNA oligonucleotide microarrays. The first approach is to use a single array to detect gene expression in multiple species, while the second approach is to use species-specific arrays. Using a single array is only applicable to closely related species and is subject to biases caused by interspecific sequence differences (Hsieh et al. 2003; Preuss et al. 2004; Gilad et al. 2005). Using multiple species-specific arrays is applicable to any species pairs, but we found that the expression divergence between species is substantially overestimated. This overestimation results from the large variation in sensitivity among different probe sets. Thus, precise measurement of expression divergence between species is still a challenging task. cDNA microarrays have also been used to assess the expression divergence between

species (Ranz et al. 2003; Renn, Aubin-Horth, and Hofmann 2004). Our method of analysis (e.g., use of *RA* instead of *S*) applies to cDNA array data as well. We think that advances in both microarray technology and statistical methodology are needed to better characterize the evolution of gene expression, which is central to our understanding of the mechanism of biological evolution (Rodriguez-Trelles, Tarrio, and Ayala 2005).

1.5 ACKNOWLEDGMENTS

We thank Itai Yanai for providing details of their published microarray analysis, James McDonald for technical assistance, Steve Qin and Xionglei He for discussion, and Wendy Grus and Soochin Cho for valuable comments on an earlier version of the manuscript. This study was supported by research grants from National Institutes of Health and University of Michigan to J.Z.

Figure 1.1 Expression-profile divergences of orthologous genes and randomly paired genes from humans and mice. A total of 10,607 orthologous gene pairs and 10,607 random gene pairs are analyzed. Expression divergence is measured by **(a)** Pearson's correlation coefficient *r* and **(b)** Euclidean distance *d*. Both measurements show that the expression-profile divergence of human-mouse orthologs is significantly lower than that of random gene pairs ($P<10^{-280}$ for *r* and $P<10^{-151}$ for *d*; Mann-Whitney U test).

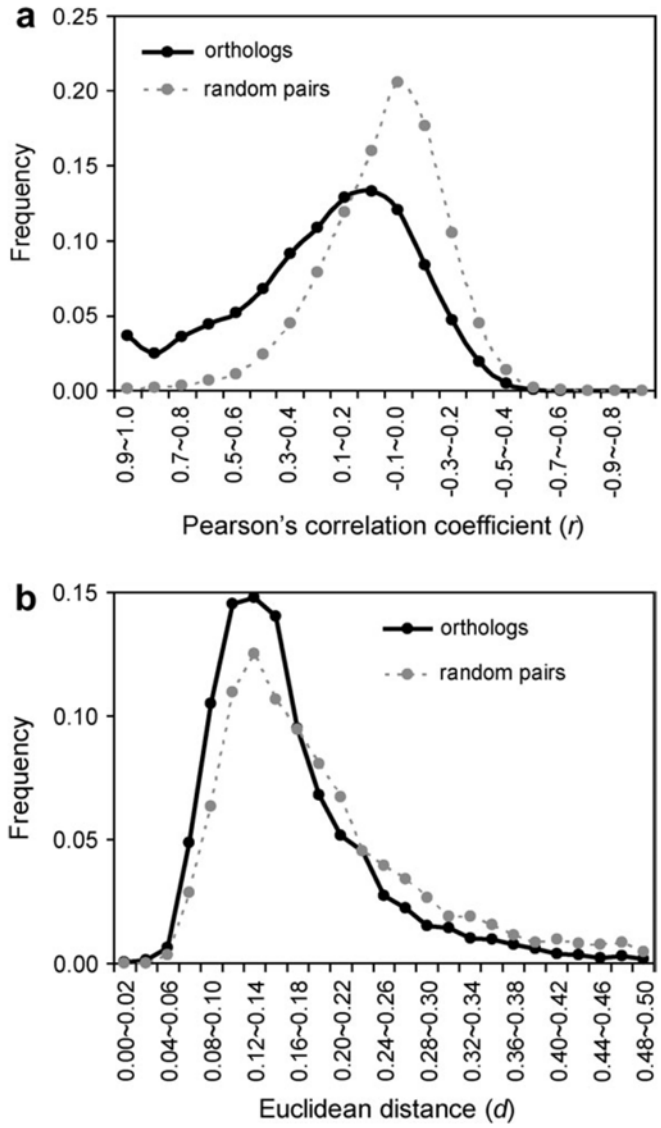


Figure 1.2 Expression profiles of the *RUTBC1* gene in humans and mice measured by the oligonucleotide microarrays. Two probe sets (#1: 212319_at; #2: 36129_at) were used on the human chip and two (#1: gnf1m22384_at; #2: gnf1m28735_at) on the mouse chip. Note that none of them are so-called "suboptimal" probe sets. (**a**) Signal intensity values. (**b**) Relative abundance values. The expression-profile similarity measured by *r* is 0.744 and 0.229, respectively, between the two probe sets for the human gene and between the two probe sets for the mouse gene. When the inter-specific divergence is estimated by comparing the expression profile measured by a human probe set and that measured by a mouse probe set, four possible *r* values can be obtained with equal probability: 0.484, 0.449, 0.495 and 0.164. The mean of the four *r* values is 0.398.

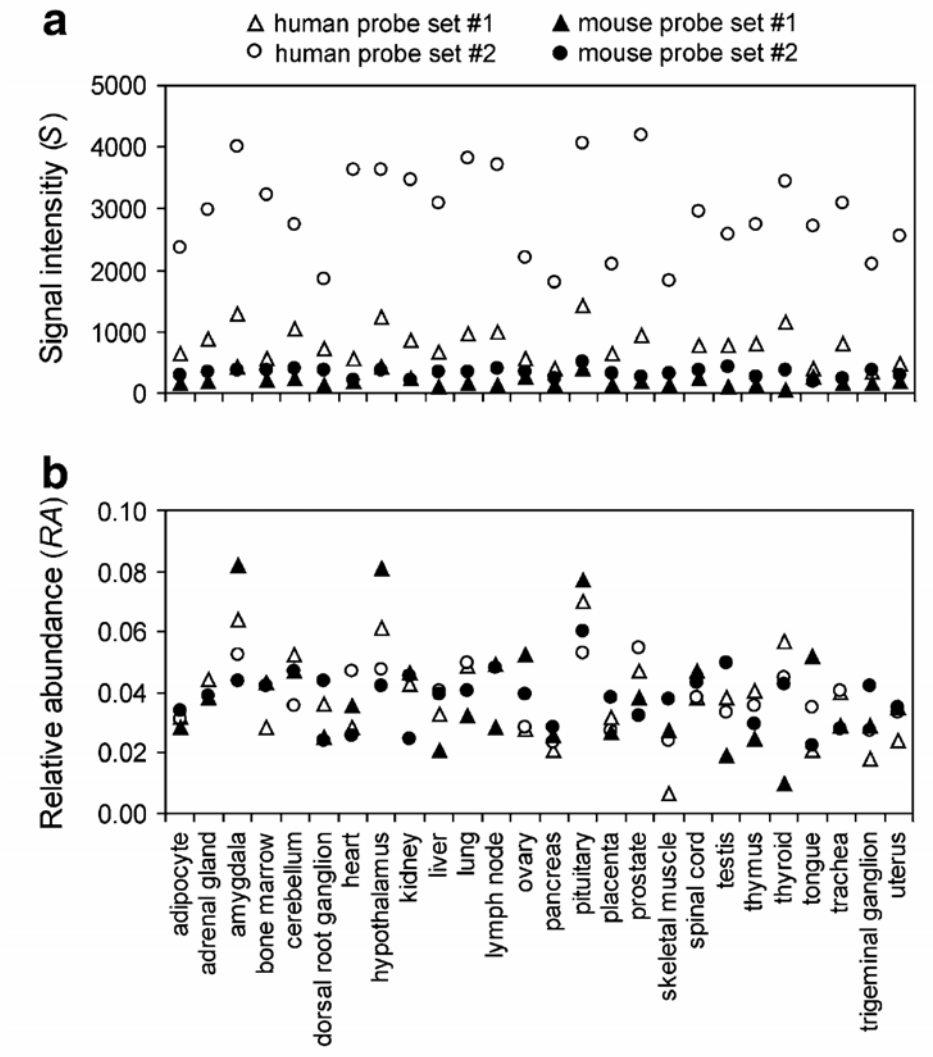


Figure 1.3 Expression-profile divergences within and between species. Presented are the results obtained from 3762 human genes and 1385 mouse genes for which multiple probe sets per gene are available. Two probe sets for each gene are randomly picked on a microarray to measure the expression-profile divergence within species (i.e., variation due to estimation errors). Two probe sets, one from the human gene and other from the orthologous mouse gene, are picked to measure the between-species divergence. The experiment is repeated 5 times. Within-species divergences are denoted in blue for human and red for mouse, while the between-species divergences are represented by green lines. The expression-profile divergence is measured by **(a)** Pearson's correlation coefficient r and **(b)** Euclidean distances d. Both measures show that within-species expression divergences are significantly lower than between-species divergences (when r is used, $P<10^{-115}$ for human and $P<10^{-5}$ for mouse; when d is used, $P<10^{-100}$ for human and $P<10^{-35}$ for mouse; Mann-Whitney U test).

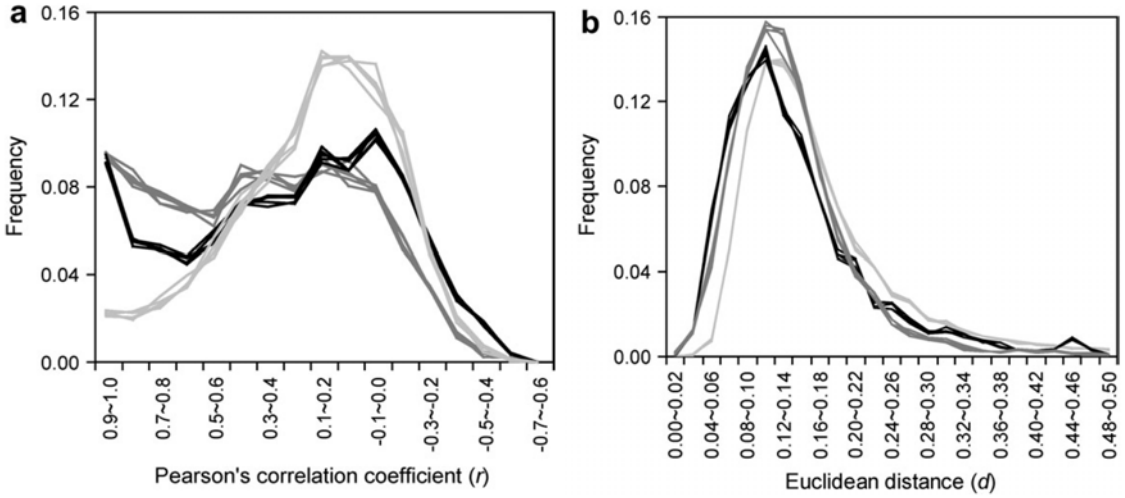


Figure 1.4 Net distances (*D*) of expressional profiles between human and mouse orthologs and Euclidean distances (*d*) of random human-mouse gene pairs. The distribution of the random pairs represents the neutral expectation of expressional divergences. The black area left to the vertical dashed line ($d_{5\%}$=0.0089) shows the 5% smallest *d* values. 83.9% of 4564 human-mouse orthologous genes have *D* smaller than $d_{5\%}$, suggesting that the detectable expression-profile divergence of 83.9% of genes is lower than the neutral expectation at the 5% significance level.

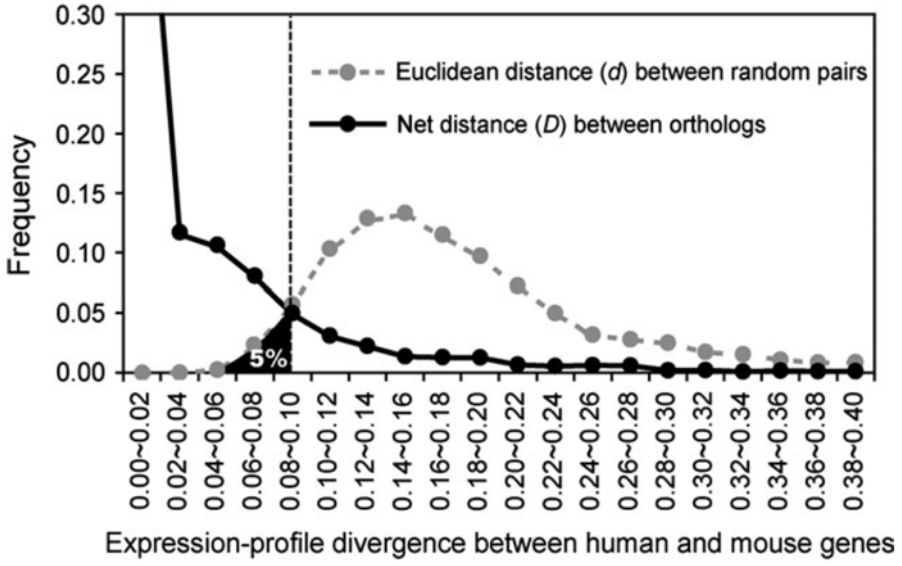


Figure 1.5 Dendrograms of 26 human and 26 mouse tissues based on: (**a**) 1-Pearson's correlation coefficient *r* and (**b**) Euclidian distance *d* of tissues. The consensus trees from 1000 bootstrap trees are presented, with the support values (bootstrap percentages) shown on branches. Interior branches with support values lower than 40 are collapsed.

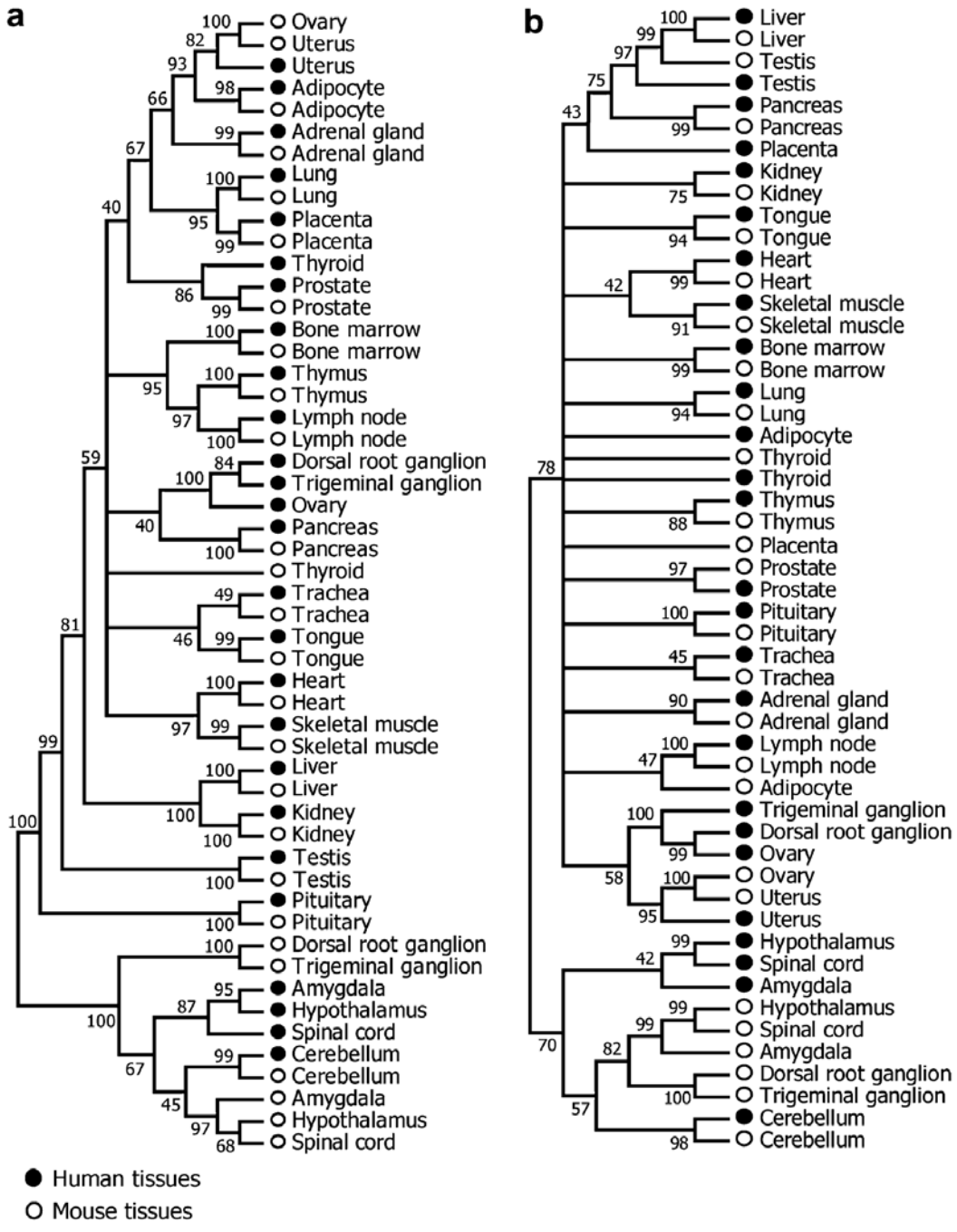
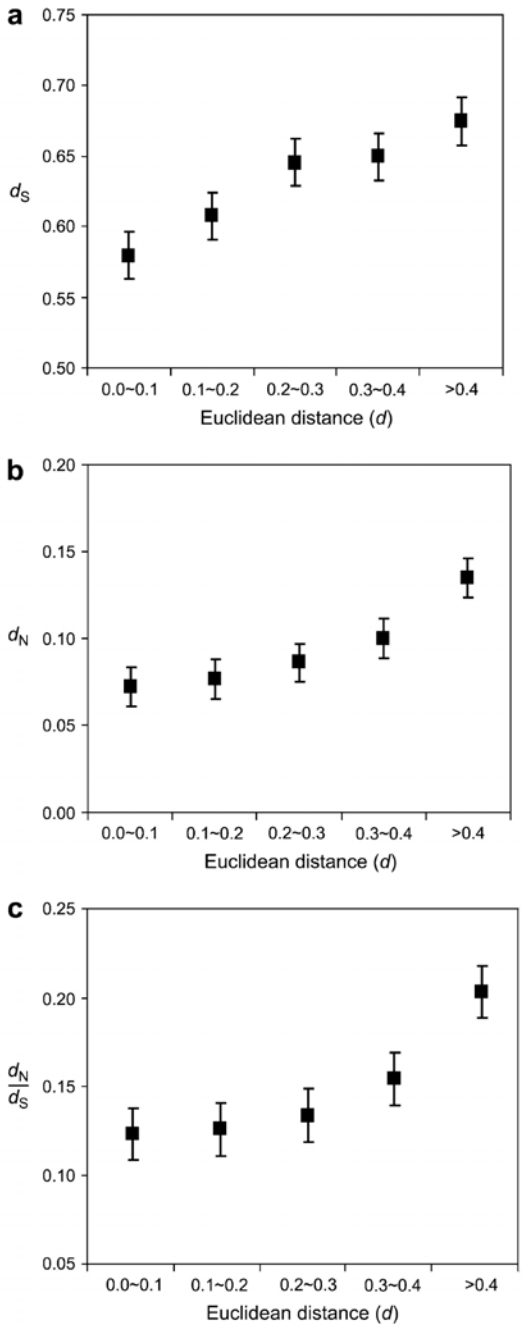


Figure 1.6 Correlation between expression-profile divergence and coding sequence divergence. Expression-profile divergence between species, measured by the Euclidean distance *d*, is positively correlated with the coding sequence divergence measured by **(a)** synonymous distance d_S, **(b)** nonsynonymous distance d_N, and (**c**) d_N/d_S. Averaged (± standard error) d_S, d_N or d_N/d_S are shown for each group of orthologs categorized by *d*. The number of human-mouse orthologous gene pairs per category is 1689, 5739, 1670, 544, and 494, respectively, for the five categories.



1.6 LITERATURE CITED

Affymetrix. 2001. Technical Note: New statistical algorithms for monitoring gene expression on GeneChip® probe arrays.

Binder, H., T. Kirsten, I. L. Hofacker, P. F. Stadler, and M. Loeffler. 2004a. Interactions in oligonucleotide hybrid duplexes on microarrays. J. Phys. Chem. B 108:18015-18025.

Binder, H., T. Kirsten, M. Loeffler, and P. F. Stadler. 2004b. Sensitivity of microarray oligonucleotide probes: variability and effect of base composition. J Phys Chem B 108:18003-18014.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D. H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S. R. Williams, K. Moon, T. Burcham, M. Pallas, R. B. DuBridge, J. Kirchner, K. Fearon, J. Mao, and K. Corcoran. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630-634.

Castillo-Davis, C. I., D. L. Hartl, and G. Achaz. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530-1536.

Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc Natl Acad Sci U S A 97:12369-12374.

Conant, G. C., and A. Wagner. 2004. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci 271:89-96.

de Bivort, B., S. Huang, and Y. Bar-Yam. 2004. Dynamics of cellular level function and regulation derived from murine expression array data. Proc Natl Acad Sci U S A 101:17687-17692.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet 37:544-548.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. Science 296:340-343.

Gilad, Y., S. A. Rifkin, P. Bertone, M. Gerstein, and K. P. White. 2005. Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res 15:674-680.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18:609-613.

Haberer, G., T. Hindemitt, B. C. Meyers, and K. F. Mayer. 2004. Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. Plant Physiol 136:3009-3022.

Hinds, H. L., C. T. Ashley, J. S. Sutcliffe, D. L. Nelson, S. T. Warren, D. E. Housman, and M. Schalling. 1993. Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. Nat Genet 3:36-43.

Hsieh, W. P., T. M. Chu, R. D. Wolfinger, and G. Gibson. 2003. Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics 165:747-757.

Hubbell, E., W. M. Liu, and R. Mei. 2002. Robust estimators for expression analysis. Bioinformatics 18:1585-1592.

Huminiecki, L., and K. H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res 14:1870-1879.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249-264.

Jordan, I. K., L. Marino-Ramirez, and E. V. Koonin. 2005. Evolutionary significance of gene expression divergence. Gene 345:119-126.

Jordan, I. K., L. Marino-Ramirez, Y. I. Wolf, and E. V. Koonin. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol 21:2058-2070.

Kasprzyk, A., D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res 14:160-169.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol 2:682-689.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309:1850-4.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, MA.

Kumar, S., K. Tamura, and M. Nei. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5:150-163.

Leung, Y. F., and D. Cavalieri. 2003. Fundamentals of cDNA microarray data analysis. Trends Genet 19:649-659.

Li, W.-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13:1638-1645.

Murtagh, F. D. 1985. "Multidimensional Clustering Algorithms" in CompStat Lectures 4. Physica-Verlag,Vienna.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261-266.

Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927-931.

Preuss, T. M., M. Caceres, M. C. Oldham, and D. H. Geschwind. 2004. Human brain evolution: insights from microarrays. Nat Rev Genet 5:850-860.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300:1742-1745.

Renn, S. C., N. Aubin-Horth, and H. A. Hofmann. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC Genomics 5:42.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet 33:138-144.

Rocha, E. P., and A. Danchin. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol 21:108-116.

Rodriguez-Trelles, F., R. Tarrio, and F. J. Ayala. 2005. Is ectopic expression caused by deregulatory mutations or due to gene-regulation leaks with evolutionary potential? Bioessays 27:592-601.

Ruan, Y., P. Le Ber, H. H. Ng, and E. T. Liu. 2004. Interrogating the transcriptome. Trends Biotechnol 22:23-30.

Slonim, D., P. Tamayo, J. P. Mesirov, T. Golub, and E. Lander. 2000. Class prediction and discoveryusing gene expression data. Pp. 263-272. Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB 2000). Universal Academy Press.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A 99:4465-4470.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101:6062-6067.

Townsend, J. P., D. Cavalieri, and D. L. Hartl. 2003. Population genetic variation in genome-wide gene expression. Mol Biol Evol 20:955-963.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler. 1995. Serial analysis of gene expression. Science 270:484-487.

Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc Natl Acad Sci U S A 97:6579-6584.

Wen, X., S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. 1998. Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci U S A 95:334-339.

Wyckoff, G. J., C. M. Malcom, E. J. Vallender, and B. T. Lahn. 2005. A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet 21:381-385.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS 8:15-24.

Yang, J., A. I. Su, and W.-H. Li. 2005. Gene Expression Evolves Faster in Narrowly than in Broadly Expressed Mammalian Genes. Mol Biol Evol 22:2113-2118.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Zhang, J., and X. He. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol 22:1147-1155.

CHAPTER 2

DIFFERENTIAL EVOLUTIONARY RATES OF MAMMALIAN GENE EXPRESSION

2.1 ABSTRACT

Evolutionary rates provide important information about the pattern and mechanism of evolution. Although the rate of gene sequence evolution has been well studied, the rate of gene expression evolution is poorly understood. In particular, it is unclear whether the gene expression level and tissue-specificity influence the divergence of expression profiles between orthologous genes. Here we address this question using a microarray dataset comprising the expression signals of 10,607 pairs of orthologous human and mouse genes from over 60 tissues per species. We show that the level of gene expression and the degree of tissue-specificity are generally conserved between the human and mouse orthologs. The rate of gene expression-profile change during evolution is negatively correlated with the level of gene expression, measured by either the average or the highest level among all tissues examined. This is analogous to the observation that the rate of gene (or protein) sequence evolution is negatively correlated with the gene expression level. The impacts of the degree of tissue-specificity on the evolutionary rate of gene sequence and that of expression-profile, however, are opposite. Highly tissue-specific genes tend to evolve rapidly at the gene sequence level, but slowly at the expression-profile level. Thus, different forces and selective constraints must underlie the evolution of gene sequence and that of gene expression.

2.2 INTRODUCTION

It has been proposed that evolutionary changes of morphology and development are more often due to alterations of gene expressions than protein sequences (King and Wilson 1975; Carroll 2005). However, compared to our knowledge of gene and protein sequence evolution (Li 1997; Nei and Kumar 2000), genome-wide patterns of gene expression evolution (Cavalieri, Townsend, and Hartl 2000; Enard et al. 2002; Oleksiak, Churchill, and Crawford 2002; Ranz et al. 2003; Rifkin, Kim, and White 2003) are poorly understood, except for the divergences of duplicate genes (Gu et al. 2002; Makova and Li 2003; Gu et al. 2004; Huminiecki and Wolfe 2004; Gu, Zhang, and Huang 2005; He and Zhang 2005). The advancement of high-throughput technologies for characterizing the expressions of thousands of genes simultaneously and the subsequent availability of microarray expression data from multiple species open the door for searching for general principles governing gene expression evolution. Two recent studies suggested that expression evolution is largely neutral, with little influences of either positive or purifying selection (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004). However, subsequent experimental studies and computational analysis using microarray-based expression data suggested that the expression evolution of most genes is subject of purifying selection (Denver et al. 2005; Jordan, Marino-Ramirez, and Koonin 2005; Rifkin et al. 2005; Liao and Zhang 2006). For example, in Chapter 1 we estimated that 84% of mammalian genes have significantly lower expression divergence than expected under complete neutrality. These findings raise the question about the determinants of the level of purifying selection on gene expression.

Evolutionary changes of gene expression can be studied from two aspects: (1) changes of gene expression level in a given tissue or under a certain condition and (2)

changes of gene expression-profile across spatial, temporal, or environmental dimensions. The first aspect has been studied more than the second (e.g., Ranz et al. 2003; Khaitovich et al. 2004). Therefore, we focus on the second aspect in this work. Specifically, we examine expression-profile evolution of mammalian genes across tissues by comparing human and mouse orthologs. Pearson's correlation coefficient (r) is used to measure the expression-profile similarity between a pair of orthologous genes. Because all human-mouse orthologs have diverged for the same amount of time, one can use r to compare the relative rates of expression-profile evolution among genes. That is, higher r indicates a lower rate of evolution, whereas lower r indicates a higher rate of evolution.

Here we consider two potential determinants of the rate of gene expression-profile evolution: expression level and tissue-specificity. These two factors were previously shown to be major determinants of the rate of gene (or protein) sequence evolution (Hastings 1996; Duret and Mouchiroud 2000; Pal, Papp, and Hurst 2001; Subramanian and Kumar 2004; Zhang and Li 2004; Zhang and He 2005) and our analysis would answer whether gene sequence evolution and expression-profile evolution are governed by the same rules. Furthermore, a recent study showed that the expression divergence between a pair of human-mouse orthologs is negatively correlated with the number of tissues in which the gene is expressed (Yang, Su, and Li 2005). This finding is puzzling, because highly specific tissue expression of a gene indicates that the gene performs a tissue-specific function (e.g., chemoreception or immunity) and it would be unlikely for such a highly specialized gene to perform functions useful to other tissues in a different species. Here we analyze the Gene Atlas V2 microarray dataset (Su et al. 2004), which includes the expression information of 10,607 human and mouse orthologous genes in over 60 tissues. Our analysis indicates that

the evolutionary rate of gene expression-profile is negatively correlated with the level of expression and the degree of tissue-specificity.

2.3 MATERIALS AND METHODS

2.3.1 Gene expression data

We used the human and mouse gene expression information from the Gene Atlas V2 dataset (http://symatlas.gnf.org/), which contains the expression data obtained by hybridization of RNAs from 73 human non-pathogenic tissues and 61 mouse tissues onto the Affymetrix microarray chips (human: U133A/GNF1H; mouse: GNF1M) designed according to the annotated human and mouse genome sequences (Su et al. 2004). A gene is represented on a chip by at least one probe set, each of which comprises several pairs of probes that overlap in their nucleotide sequences. To assign the probe sets to the current annotated version of Ensembl human and mouse genes, we aligned sequences of each probe set to the Ensembl cDNA sequences (human: Homo_sapiens.NCBI35.feb.cdna.fa; mouse: Mus_musculus.NCBIM33.feb.cdna.fa; http://www.ensembl.org/) using BLASTn (http://www.ncbi.nlm.nih.gov/blast/) and kept those probe sets in which all matching probes perfectly matched to the same Ensembl gene. 25,368 probe sets (75.3%) in the human chip corresponding to 16,456 genes and 18,005 probe sets (49.8%) in the mouse chip corresponding to 15,835 genes were retained for further analysis. The expression level detected by each probe set was obtained as the signal intensity (*S*) computed from either MAS 5.0 algorithm (MAS5) (Hubbell, Liu, and Mei 2002) or GC-content-adjusted robust multi-array algorithm (GC-RMA) (Wu et al. 2004). The Gene Atlas V2 dataset derived from

GC-RMA algorithm was downloaded from GNF Genome Informatics Applications & Datasets (http://wombat.gnf.org). The *S* values were averaged among replicates. Because the results from MAS5 and GC-RMA are similar, we present the findings obtained from MAS5 unless otherwise noted.

2.3.2 Tissue-specificity of gene expression

We used Tissue Specificity Index τ (Yanai et al. 2005) to measure the tissue-specificity of a human or mouse gene. The τ of human gene *i* is defined by

$$\tau_{\mathrm{H}} = \frac{\sum_{j=1}^{n_{\mathrm{H}}} \left(1 - \left[\frac{\log_2 S_{\mathrm{H}}(i,j)}{\log_2 S_{\mathrm{H}}(i,\max)}\right]\right)}{n_{\mathrm{H}} - 1}, \qquad (1)$$

where n_{H} is the number of human tissues examined and $S_{\mathrm{H}}(i, \max)$ is the highest expression signal of gene *i* across the n_{H} tissues. To minimize the influence of noise from low intensities, we arbitrarily let $S_{\mathrm{H}}(i, j)$ be 100 if it is lower than 100. Note that this strategy of reducing the effect of noise is used only in computing τ. When a gene has more than one probe set on the chip, we compute τ by averaging the τ values derived from the different probe sets. The τ value ranges from 0 to 1, with higher values indicating higher variations in expressional level across tissues, or higher tissue specificities. If a gene has expression in only one tissue, τ approaches 1. In contrast, if a gene is equally expressed in all tissues, $\tau = 0$.

2.3.3 Human-mouse orthologs

The homology information of human and mouse genes was obtained from Ensembl EnsMart (http://www.ensembl.org/Multi/martview) (Kasprzyk et al. 2004). There are several

annotated homology relationships between human and mouse genes by Ensembl. We only considered those pairs of genes annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BLASTZ searches) to be orthologous. We found that 10,607 pairs of human-mouse orthologs have expression data. Affymetrix probes with name suffixes _x_at and _s_at were thought to be prone to cross-hybridization, compared to other probes (Affymetrix Technical Support, Data Analysis Fundamentals, Appendix B; http://www.affymetrix.com/), and have been considered "suboptimal" (Yang, Su, and Li 2005). But our recent analysis showed that the quality of these probes is not worse than other probes (Liao and Zhang 2006). We therefore considered all probe sets equally.

The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse orthologs were retrieved from Ensembl EnsMart. In this database, d_S and d_N were estimated by the maximum likelihood method using the PAML package (Yang 1997).

2.3.4 Expression-profile similarity between orthologous genes

To measure the similarity in expression-profile between human and mouse orthologs, we analyzed 26 common tissues of the two species included in the dataset (Su et al. 2004). These 26 tissues are adipocyte, adrenal gland, amygdala, bone marrow, cerebellum, dorsal root ganglion, heart, hypothalamus, kidney, liver, lung, lymph node, ovary, pancreas, pituitary, placenta, prostate, skeletal muscle, spinal cord, testis, thymus, thyroid, tongue, trachea, trigeminal ganglion, and uterus. Mouse lower spinal cord was used as the

homologous tissue of human spinal cord. We measured the expression-profile similarity between a pair of orthologous genes by Pearson's correlation coefficient *r*, defined as

$$r = \frac{\sum_{j=1}^{n}[S_{\mathrm{H}}(i,j)S_{\mathrm{M}}(i,j)] - [\sum_{j=1}^{n} S_{\mathrm{H}}(i,j)][\sum_{j=1}^{n} S_{\mathrm{M}}(i,j)]/n}{\sqrt{\sum_{j=1}^{n}[S_{\mathrm{H}}(i,j)]^2 - [\sum_{j=1}^{n} S_{\mathrm{H}}(i,j)]^2/n}\sqrt{\sum_{j=1}^{n}[S_{\mathrm{M}}(i,j)]^2 - [\sum_{j=1}^{n} S_{\mathrm{M}}(i,j)]^2/n}}. \quad (2)$$

Here, n =26 is the number of common tissues considered, H indicates human, and M indicates mouse. $S_{\mathrm{H}}(i, j)$ and $S_{\mathrm{M}}(i, j)$ are the expression signal intensity of gene i in human tissue j and mouse tissue j, respectively. A high r indicates a high similarity in expression-profile between the orthologs and a low rate of expression-profile evolution. Note that in our previous study (Liao and Zhang 2006), the relative abundance of mRNA across tissues (the signal of one tissue relative to the total signal of all tissues) was used to compute r. In fact, using either relative abundance or S gives exactly the same r value. To compare our results with those of Yang, Su, and Li (2005), we also used the Expression Conservation Index (*ECI*) that they developed. The *ECI* between a pair of human-mouse orthologs is

$$ECI = \frac{N_{\mathrm{HM}} + 0.5}{(N_{\mathrm{H}} + N_{\mathrm{M}})/2 + 0.5}, \quad (3)$$

where N_{H} and N_{M} are the numbers of human and mouse tissues in which the gene is expressed, respectively, and N_{HM} is the number of tissues in which the gene is expressed in both species. According to Yang, Su, and Li (2005), a gene is considered to be expressed in a tissue if $S \geq 200$ for the tissue. *ECI* varies from 0 to 1, with higher values indicating higher similarity between expression profiles. When a gene is represented by more than one probe set on a microarray chip, r and *ECI* are computed by averaging the values obtained from all possible combinations of a human probe set and a mouse probe set of the gene.

2.4 RESULTS AND DISCUSSIONS

2.4.1 Choice of parameters used in this study

The transcriptome data analyzed in the present study were obtained from oligonucleotide microarray experiments. When quantifying tissue-specificity of a gene or comparing expression-profile similarity between a pair of orthologs, it is important to consider properties of microarray data.

Tissue-specificity of gene expression measures the degree of differential expression across tissues. It is expected that a gene with higher tissue-specificity tends to have lower expression breadth (*B*), which is the proportion of tissues in which the gene is expressed. In microarray data analysis, the number of tissues (*N*) in which a gene is expressed is usually determined by an arbitrary cutoff of the signal intensity *S* (Su et al. 2002; Vinogradov 2004; Yang, Su, and Li 2005). Similar definitions of tissue-specificity have also been used in studies based on SAGE (serial analysis of gene expression) or EST (expression sequence tag) data (Duret and Mouchiroud 2000; Ponger, Duret, and Mouchiroud 2001; Lercher, Urrutia, and Hurst 2002; Subramanian and Kumar 2004). However, there are several problems with the approach of applying a cutoff in defining whether a gene is expressed in a tissue. First, the number of mRNA molecules of a gene in a given tissue is a continuous figure; expression should not be characterized as absent or present. Second, the expression level required for a gene to be functional presumably varies substantively among genes; it is unreasonable to use a single cutoff for all genes in all tissues. Third, expression breadth actually measures the tissue restriction of expression, but ignores quantitative variations in expression among many tissues (Schug et al. 2005). Fourth, the *S* value in microarray data is not only determined by the quantity of the target mRNA, but also by the probe-target affinity and the algorithm of

raw-data processing. In other words, two genes with the same *S* values do not necessarily have the same mRNA concentration. Although a recent study (Khaitovich et al. 2005) used the Affymetrix detection *p*-value instead of the cutoff value of *S* to determine the expression status of a gene in a given tissue, several of the above problems cannot be avoided. Because of these problems with the cutoff-based expression breadth (*B*), we use Tissue Specificity Index (τ) to measure tissue-specificity. Use of the parameter τ can avoid the aforementioned problems.

Another potential measure of tissue-specificity is the coefficient of variation (*CV*) of expression across tissues. *CV* is defined as the standard deviation of a random variable divided by its mean. The *CV* value for human gene *i* can be computed by the standard deviation of $\log_2 S_H(i, j)$ among the 73 human tissues considered divided by the average $\log_2 S_H(i, j)$ of the 73 tissues. A high *CV* indicates a great variation in gene expression among tissues, implying tissue-specificity. Because τ and *CV* are highly correlated (Spearman's rank correlation coefficient = 0.693, $P<10^{-300}$; Pearson's correlation coefficient = 0.690, $P<10^{-300}$; see Figure A.4), we will only use τ in this study.

We use Pearson's correlation *r* to measure the expression-profile similarity (conservation) between a pair of orthologous genes. It was claimed by Yang, Su and Li (2005) that compared to *r*, *ECI* is a more appropriate measure. However, our results suggest that *r* is better than *ECI* in quantifying expression-profile similarity (see below). For instance, unlike *ECI*, using *r* avoids the use of cutoff-based method in defining *N*.

2.4.2 Conservation of gene expression level and tissue-specificity during evolution

We examine whether the level of gene expression and the degree of tissue-specificity are similar between human and mouse orthologous genes. If gene expression evolution is not selectively constrained, as suggested earlier (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), no such similarity is expected (Jordan, Marino-Ramirez, and Koonin 2005), because of the long divergence time between the two species (Springer et al. 2003; Murphy, Pevzner, and O'Brien 2004) and the rapid pace with which gene expression can change during evolution (Gu et al. 2002). However, we found a strong positive correlation in both mean expression level (Figure 2.1a; Spearman's rank correlation coefficient = 0.392, $P < 10^{-300}$) and tissue-specificity (Figure 2.1b; Spearman's rank correlation coefficient = 0.296, $P < 10^{-212}$) between human and mouse orthologs. Note that the mean expression levels are calculated from averaging *S* values of 73 normal human tissues or 61 mouse tissues. Similar results were obtained when only the 26 common tissues between humans and mice were considered (Spearman's rank correlation coefficient = 0.384, $P < 10^{-300}$, for mean expression level; Spearman's rank correlation coefficient = 0.335, $P < 10^{-276}$, for tissue-specificity). It is interesting to note that although the type of microarray data we analyzed were reported to be noisy (Hill et al. 2001; Irizarry et al. 2003) and probe sets of orthologous genes often have different hybridization behaviors (Liao and Zhang 2006), significant similarities in expression level and tissue-specificity are still apparent between human and mouse orthologs, strongly suggesting the evolutionary conservation of gene expression. Our result regarding the conservation of gene expression level is consistent with that of Jordan, Marino-Ramirez, and Koonin (2005).

Previous studies showed that gene expression level and expression breadth are strongly and positively correlated (Lercher, Urrutia, and Hurst 2002; Vinogradov 2004).

This is not unexpected, as expression breadth is determined by the expression-signal cutoff used. However, in the present study, virtually no correlation is found between expression level and tissue-specificity τ. For example, in humans, Spearman's rank correlation coefficient between τ and mean S is -0.007 ($P = 0.481$). Because the correlations we report in the following two sections are much higher and very significant, it is appropriate to assume that τ and S are uncorrelated.

2.4.3 Highly expressed genes have low rates of expression-profile evolution

The phenomenon that highly expressed genes have lower substitution rates than lowly expressed genes in coding-sequences has been reported in bacteria (Rocha and Danchin 2004), unicellular eukaryotes (Pal, Papp, and Hurst 2001; Wall et al. 2005; Zhang and He 2005), and multicellular eukaryotes (Subramanian and Kumar 2004; Jordan, Marino-Ramirez, and Koonin 2005). This is also true in our dataset. For example, the average expression level of human genes (S_H) and the nonsynonymous nucleotide distance d_N between human and mouse orthologs are negatively correlated (Spearman's rank correlation coefficient = -0.160, $P < 10^{-58}$). We also found a weak negative correlation between S_H and the synonymous nucleotide distance d_S (Spearman's rank correlation coefficient = -0.099, $P < 10^{-23}$). S_H and d_N/d_S are also negatively correlated (Spearman's rank correlation coefficient = -0.139, $P < 10^{-44}$). These results confirm that genes of high expression are more selectively constrained in the coding-sequence than genes of low expression. Below, we examine whether highly expressed genes are also more constrained in their expression-profile evolution.

Our analysis of 10,607 human-mouse orthologs shows that highly expressed genes

have more similar expression profiles between species than lowly expressed genes (Figure 2.2 for the binned data). This is true regardless of whether the expression level is measured by the average *S* over all tissues (Figure 2.2a: human; Figure 2.2b: mouse) or by the maximum *S* (Figure 2.2c: human; Figure 2.2d: mouse) among 73 human or 61 mouse tissues examined. For the unbinned original data, the positive correlation between profile similarity and expression level is also strong (rank correlation coefficient: 0.17 to 0.37) (Figure 2.2 legend). Because the expression-profile similarities are derived from the 26 tissues common to the human and mouse, we also conducted the correlation analysis using average *S* and maximum *S* computed from the 26 common tissues. The results obtained (Figure A.5) are similar to those presented in Figure 2.2. Furthermore, we used the GC-RMA expression dataset and obtained similar results (Figure A.6).

It is possible that the positive correlation between gene expression level and expression-profile similarity is due to the relatively strong background noise at low expression levels, which would reduce the expression-profile similarity more for lowly expressed genes. If our result is mainly due to such a factor, the correlation between the expression level and profile similarity should be much weaker in the subset of genes with high expressions. We examined genes with average $S \geq 800$, a much greater value than that commonly thought to be significant ($S = 200$, Su et al. 2002). We found that highly expressed genes ($S \geq 800$) still show the same trend (Figure 2.2a and 2.2b), suggesting that our observation is not due to the background noise. Our results thus suggest that highly expressed genes are exposed to stronger purifying selection in both coding-sequence evolution and expression-profile evolution than lowly expressed genes.

2.4.4 Tissue-specific genes have low rates of expression-profile evolution

Previous studies showed that broadly expressed genes such as housekeeping genes have lower substitution rates in their coding-sequences than narrowly expressed genes (Hastings 1996; Duret and Mouchiroud 2000; Winter, Goodstadt, and Ponting 2004; Zhang and Li 2004). It is expected that the same trend exists between tissue-specificity τ and the rate of coding-sequence evolution. Indeed, we found weak positive correlations between τ_H and d_N (Spearman's rank correlation coefficient = 0.089, $P < 10^{-18}$), d_S (Spearman's rank correlation coefficient = 0.114, $P < 10^{-24}$), and d_N/d_S (Spearman's rank correlation coefficient = 0.060, $P < 10^{-9}$). Next, we examined the relationship between tissue-specificity and the rate of expression-profile divergence. We found that genes with higher τ tend to show higher expression-profile similarity (r) between human-mouse orthologs (see Figure 2.3 for the binned data). This correlation is strong (rank correlation coefficient of 0.34-0.38) and highly significant even for the original unbinned data (see Figure 2.3 legend). The GC-RMA expression dataset gave similar results (Figure A.7). Because the correlation between r and τ is much higher than that between τ and S, we conclude that the former correlation is not due to the latter. In other words, expression level and tissue-specificity independently influence the rate of expression-profile evolution.

Our finding of the positive correlation between r and τ appears to be opposite of what Yang, Su, and Li (2005) found. They showed that broadly expressed genes have lower rates of gene expression-profile evolution than narrowly expressed genes, which was based on the observation of a positive correlation between expression breadth (B) and the Expression Conservation Index (ECI) between human and mouse orthologs. Their results may not reflect biological reality for the following three reasons.

First, as aforementioned, they used a potentially problematic approach of applying a signal cutoff to the microarray data and defining expression breadth by counting the number of tissues in which a gene is expressed. Figure 2.4a gives an example illustrating its flaws. It is common that on a microarray chip there are more than one probe sets to represent a gene. Theoretically, different probe sets of the same gene should give similar values of τ (or B) because these different probe sets target the same mRNA. However, when the cutoff value of 200 is used for the two probe sets of human *WASPIP* gene, B is substantively different depending on which probe set is used (probe set #1: B=2/26=0.077; probe set #2: B=17/26=0.654). Fig. 4a shows that the two probe sets provide relatively consistent expression patterns except that probe set #1 has much lower affinity to the target mRNA than probe set #2. Contrary to B, similar τ values were obtained using these two probe sets (probe set #1: 0.351; probe set #2: 0.334), illustrating that τ is a better measure than B.

Second, because the number of tissues in which a gene is expressed (N) is highly dependent on the signal cutoff used and because *ECI* is computed from N, one can expect that *ECI* is also problematic. For example, in Figure 2.4a, although the two probe sets represent the same human gene (*WASPIP*) and have congruent expression patterns, the *ECI* value is low (0.250). In Figure 2.4b, although human and mouse *NEU1* genes have substantively different expression profiles, *ECI* is high (0.961). Contrary to *ECI*, Pearson's r between expression profiles seems a better index reflecting biological facts (Figure 2.4a: r = 0.849; Figure 2.4b: r = 0.288).

Finally, because both *ECI* and B are computed from N, it is expected that *ECI* and B are not independent from each other. From formula (3), we expect that human-mouse orthologs with larger N should have higher *ECI* values, because by chance they have more

opportunities to overlap in expression. To demonstrate this effect, we randomly paired human and mouse genes. As shown in Figure 2.5a, the randomly paired genes still show positive correlation between *ECI* and *B*, suggesting that the previously observed correlation in Yang, Su, and Li (2005) may not be due to true biological relationships, but rather an artifact caused by the dependence between the two parameters used. By contrast, such a correlation does not exist for randomly paired genes when we use τ to measure tissue-specificity and *r* to measure expression-profile similarity (Figure 2.5b). Yang, Su, and Li (2005) attempted to avoid the dependence between *ECI* and *B* by using different sets of tissues to compute *ECI* and *B*. They suggested that their result still holds after this consideration, as shown in their Table 1. However, they did not control for the expression level *S*. Because expression breadth *B* and mean *S* are highly correlated (Spearman's rank correlation = 0.86, $P<10^{-300}$ in our data), their observation of conservation of broadly expressed genes could be due to the fact that (i) broadly expressed genes tend to have high expression and (ii) highly expressed genes tend to be conserved (Figure 2.2). The advantage of using τ instead of *B* is that τ and *S* are uncorrelated (see above).

Previous molecular evolutionary studies have considered the differences between house-keeping and non-house-keeping genes (e.g., Zhang and Li 2004). House-keeping genes are those expressed in the majority of tissues. It is expected that house-keeping genes have lower tissue-specificity than non-house-keeping genes. If one defines human house-keeping genes by $S \geq 200$ in at least 70 of the 73 examined tissues, τ is 0.168±0.001 (mean ± standard error of mean) for the 2262 house-keeping genes, but 0.225±0.001 for the other 8345 genes, consistent with the above expectation. However, the average expression level is much higher for house-keeping genes (1351±33) than for the other genes (413±5).

Interestingly, we found that the expression-profile similarity (r) between human-mouse orthologs does not differ between house-keeping genes (0.211±0.006) and the other genes (0.215±0.003). Apparently, high expression levels and low tissue-specificities offset each other so that house-keeping genes do not differ from other genes in r. We note that although house-keeping genes tend to have low variations in expression level across tissues, the variance is not 0. Furthermore, the relative expression levels across tissues may not be important to house-keeping genes. This may explain why r is not higher for house-keeping genes than for non-house-keeping genes.

2.4.5 Similarities and differences between coding-sequence and expression-profile evolution

In this work, we used statistical correlations to identify factors that might influence the evolution of expression profiles of mammalian genes. It is important to address (i) whether two quantities are significantly correlated and (ii) how strong the correlation is. The important correlations on which our main conclusions are based range from 0.17 to 0.38. These correlations are not particularly high, though statistically highly significant. The relatively low correlations may reflect two facts. First, the evolutionary rate of gene expression-profile is determined by multiple factors, each of which may only have a small effect. Second, microarray expression data are known to be noisy, which reduces correlations. Because the evolution of gene expression-profiles is poorly understood, it is important to first identify all relevant determinants before one can evaluate their relative contributions. It is also useful to compare the magnitudes of the newly identified correlations with those of well established correlations, as will be discussed below.

By analyzing over ten thousand human-mouse orthologous gene pairs, we found that highly expressed genes have lower rates of evolution than lowly expressed genes in both coding-sequence and expression-profile (Figure 2.6). Gene expression level (*S*) is thought to be the single most important determinant of the rate of coding-sequence evolution (Drummond, Raval, and Wilke 2006). We found that the correlation (0.17) between expression-profile similarity (*r*) and *S* is slightly higher than that (0.14-0.16) between d_N (or d_N/d_S) and *S* for mammalian genes, suggesting similar importance of expression level in determining the rate of expression-profile evolution and the rate of coding-sequence evolution.

Do the similar impacts of gene expression level on coding-sequence and expression-profile evolution suggest a common evolutionary mechanism? A recent study proposed that highly expressed proteins are under stronger pressures to avoid misfolding caused by translational errors; consequently, these proteins have more rigid requirements for their sequences and are more conserved in evolution (i.e., the translational robustness hypothesis) (Drummond et al. 2005). Although this hypothesis may explain why highly expressed genes have low rates of coding-sequence evolution, it cannot explain why they also have low rates of expression-profile evolution, because there is no link between expression-profile conservation and protein misfolding. It has also been proposed that highly expressed genes are functionally more important and therefore are more conserved in their coding-sequences (Rocha and Danchin 2004). This functional importance hypothesis may explain our observations if functionally important genes are under strong purifying selection in both coding-sequences and expression profiles. However, the functional importance hypothesis was not supported in a previous study of yeasts (Drummond et al. 2005). Furthermore, in

yeasts and bacteria, only a small fraction of the strong correlation between gene expression level and d_N may be explained by their covariations with gene importance, which is measured by the fitness reduction caused by gene deletions (Rocha and Danchin 2004; Zhang and He 2005). The main reason behind the impact of gene expression level on the rate of coding-sequence evolution is still unclear. It is possible that the apparently similar influences of gene expression level on coding-sequence divergence and expression-profile divergence have different underlying causes.

We found that tissue-specificity has opposite impacts on the rate of coding-sequence evolution and the rate of expression-profile evolution. Compared with a gene with low tissue-specificity, a gene with high tissue-specificity tends to evolve faster in its coding-sequence, but slower in its expression profile (Figure 2.6). It has been suggested that there is less functional constraint on a protein sequence if it is expressed only in a small number of tissues (Duret and Mouchiroud 2000). At the same time, tissue-specific genes may be more adaptable due to fewer pleiotropic effects (Duret and Mouchiroud 2000). As a consequence, tissue-specificity and d_N become positively correlated. More detailed causal effects regarding this relationship have been discussed in Zhang and Li (2004). However, it is worth noting that the correlation between tissue-specificity (τ) and d_N is low (Spearman's rank correlation = 0.089) in our analysis. Previous studies demonstrating an impact of tissue-specificity on coding-sequence evolution were likely confounded by the influence of expression level, as expression cutoffs were used to define tissue-specificity (Duret and Mouchiroud 2000; Zhang and Li 2004). In the present study, however, the impact of tissue-specificity can be clearly separated, as τ is uncorrelated with expression level.

The correlation between expression-profile similarity and τ ranges from 0.34 to 0.38, indicating that the impact of tissue-specificity on expression-profile evolution is much greater than that on coding-sequence evolution. Given the large estimation error of expression-profile similarity caused by microarray technologies (Liao and Zhang 2006), the high correlation observed prompts us to believe that tissue-specificity is one of the most important determinants of the evolutionary rate of gene expression-profile in mammals. Why do highly tissue-specific genes have a low rate of expression-profile evolution? It is possible that for a tissue-specific gene, its function is highly specialized for the tissue(s) where it is expressed. Expression of the gene in a different tissue would make the protein physiologically useless or even detrimental. In contrast, non-tissue-specific genes may be more tolerant to changes of expression level in various tissues, thus having relatively high rates of expression-profile evolution. Taken together, expression-profile evolution and coding-sequence evolution appear to be governed by different principles.

A recent study based on human-chimpanzee comparisons suggested that the evolutionary rate of the expression level of a gene is positively correlated with the evolutionary rate of its coding-sequence (Khaitovich et al. 2005). However, it is unclear whether the evolutionary rate of expression-profile is correlated with that of coding-sequence (Figure 2.6). Several studies using human-mouse orthologs do not find such a correlation (Jordan et al. 2004; Yanai, Graur, and Ophir 2004; Jordan, Marino-Ramirez, and Koonin 2005). Our previous study revealed a weak positive correlation between these two quantities when the Euclidean distance was used to measure the profile similarity of human-mouse orthologs (Liao and Zhang 2006). However, such a correlation was not observed when Pearson's *r* was used to measure the profile similarity. Figure 2.6 illustrates that these

ambiguous results might be related to the different effects of the expression level and tissue-specificity on the evolutionary rate of coding-sequence and that of expression profile.

It should be emphasized that genome-wide analysis of gene expression evolution has just begun and most studies have focused on identifying correlations. When a higher quantity and quality of data become available, the underlying causes of the identified correlations and the relative contributions of various factors may be examined. We also want to stress that the impacts of expression level and tissue-specificity on the evolutionary rate of expression profile that we report in this work should be confirmed in other datasets and other species. Unlike the study of gene/protein sequence evolution, in which various evolutionary distances have been developed (Li 1997; Nei and Kumar 2000), the study of expression-profile divergence still lacks a good distance measure. All the distances so far introduced (*r*, Euclidian distance, and *ECI*) only measure the relative divergence of expression profiles, but tell nothing about the number of genetic changes that are responsible for the expression divergence. Understanding the molecular genetic mechanisms of expression regulation will facilitate the development of such distance measures, which will in turn help elucidate the mode and cause of expression evolution.

2.5 ACKNOWLEDGMENTS

We thank Luis Chaves, Wendy Grus, Xionglei He, Ondrej Podlaha, and two anonymous reviewers for valuable comments. This work was supported by research grants from University of Michigan and National Institutes of Health to J.Z.

Figure 2.1 Expression similarity between human-mouse orthologs in (a) mean expression level and (b) tissue-specificity. Spearman's rank correlation coefficient = 0.392 ($P < 10^{-300}$) for panel (a) and 0.296 ($P < 10^{-212}$) for panel (b). In addition, the linear regression and Pearson's correlation coefficient (R) is presented for each panel. The data include 10,607 human-mouse orthologs. The mean expression levels (S_H or S_M) and tissue-specificity (τ_H or τ_M) of the human and mouse genes are calculated from 73 human and 61 mouse normal tissues, respectively.

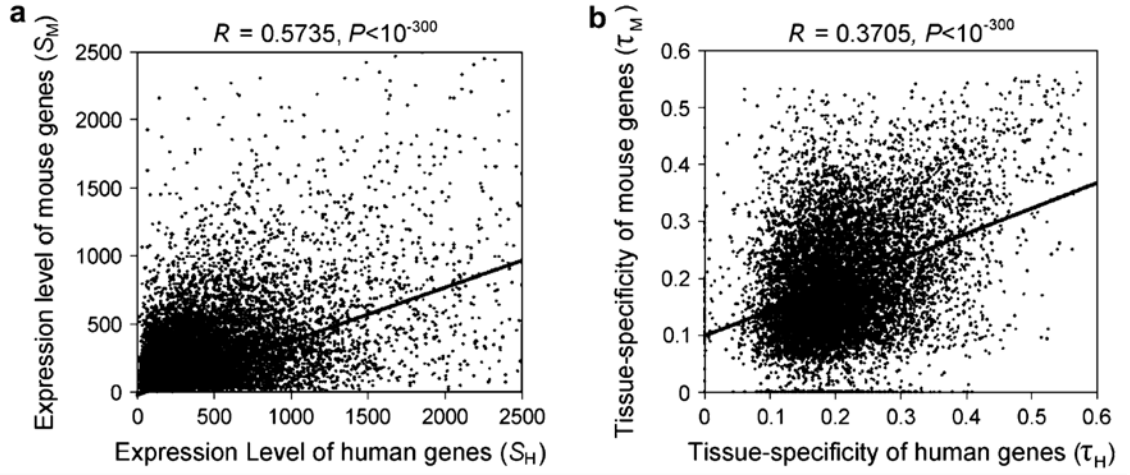


Figure 2.2 Highly expressed genes have higher expression-profile similarity between human-mouse orthologs than lowly expressed genes (MAS5 dataset). The expression level is measured by either the mean expression level or the maximum expression level across all tissues (i.e., 73 human normal tissues or 61 mouse tissue). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The data include 10,607 human-mouse orthologs. We measured the correlations using the original unbinned data. Spearman's rank correlation coefficient is (**a**) 0.172 ($P < 10^{-71}$), (**b**) 0.176 ($P < 10^{-74}$), (**c**) 0.333 ($P < 10^{-272}$), and (**d**) 0.365 ($P < 10^{-300}$), respectively. The number of gene pairs used in each bin is: **(a)** 0-200: 2517, 200-400: 2781, 400-800: 3093, 800-1600: 1576, >1600: 640; **(b)** 0-200: 4377, 200-400: 3132, 400-800: 2064, 800-1600: 768, >1600: 266; **(c)** 0-400: 909, 400-800: 1900, 800-1600: 2743, 1600-3200: 2302, 3200-6400: 1472, >6400: 1281; **(d)** 0-400: 2439, 400-800: 2507, 800-1600: 2402, 1600-3200: 1619, 3200-6400: 961, >6400: 679.

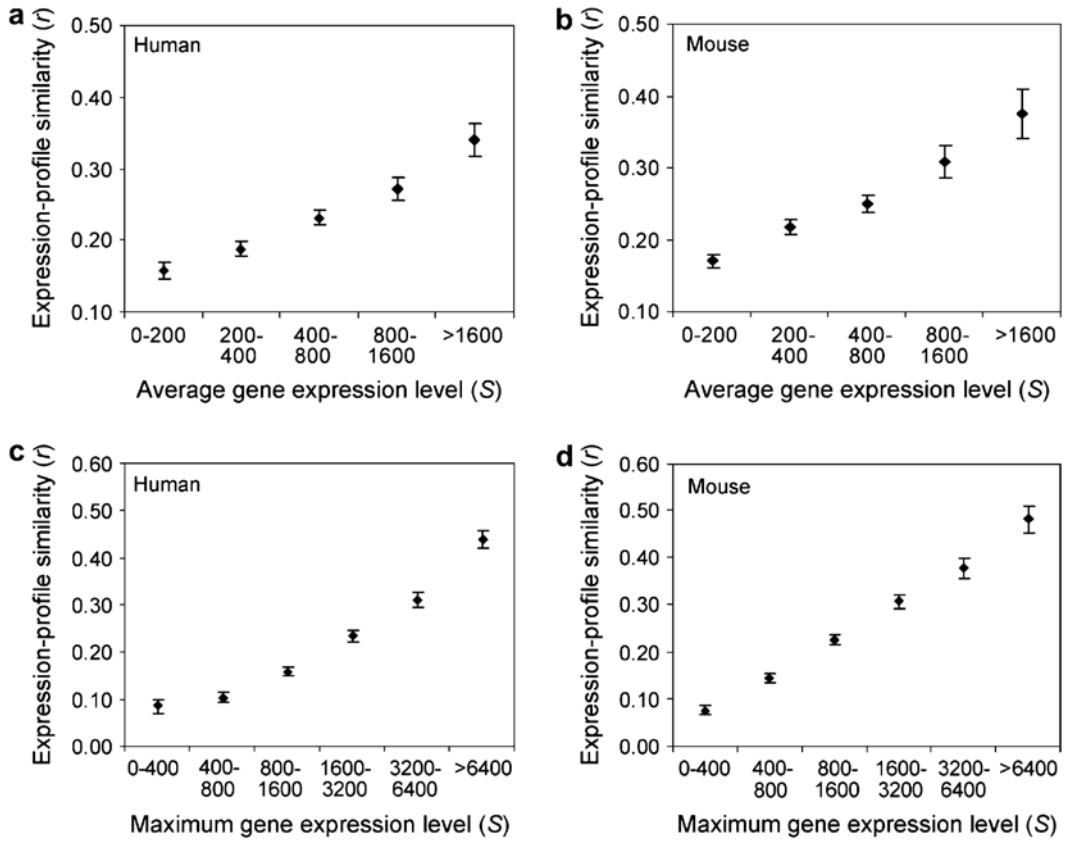


Figure 2.3 Greater expression-profile similarities between human-mouse orthologs for genes of high tissue-specificity than genes of low tissue-specificity (MAS5 dataset). Tissue-specificity is measured using all tissues (i.e., 73 human normal tissues or 61 mouse tissue). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The data include 10,607 human-mouse orthologs. We measured the correlations using the original unbinned data. Spearman's rank correlation coefficient is (**a**) 0.340 ($P < 10^{-285}$) and (**b**) 0.377 ($P < 10^{-300}$). The numbers of genes in each bin are: **(a)** 0.00-0.05: 84, 0.05-0.10: 397, 0.10-0.15: 1810, 0.15-0.20: 3146, 0.20-0.25: 2352, 0.25-0.30: 1305, 0.30-0.35: 756, 0.35-0.40: 397, >0.40: 360; **(b)** 0.00-0.05: 444, 0.05-0.10: 1184, 0.10-0.15: 2473, 0.15-0.20: 2151, 0.20-0.25: 1613, 0.25-0.30: 1117, 0.30-0.35: 740, 0.35-0.40: 444, >0.40: 441.

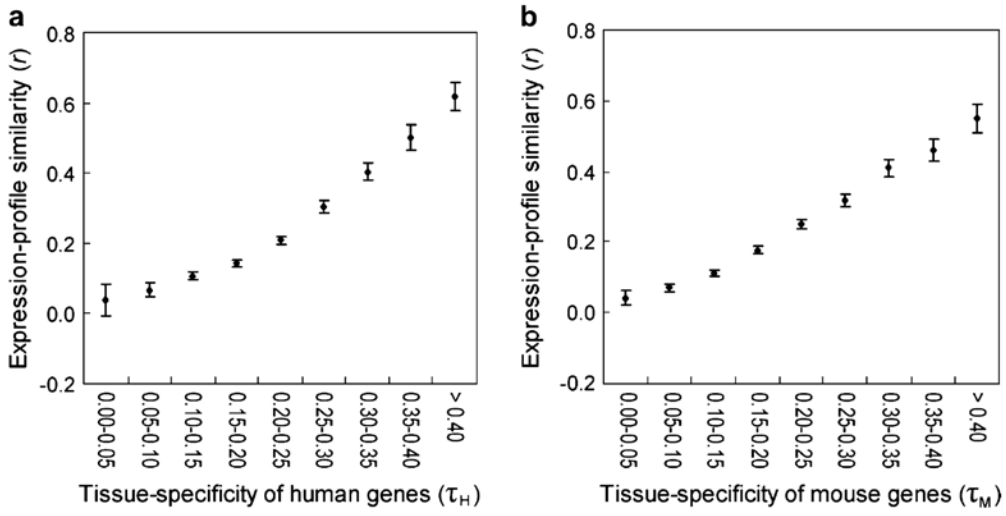


Figure 2.4 Two examples of expression profiles obtained from Gene Atlas V2. (a) Profiles of two probe sets (probe set #1: 202663_at; probe set #2: 202664_at) of human *WASPIP* gene. Expression breadth (*B*) for the probe set #1 and probe set #2 is 0.077 and 0.654, respectively. Tissue-specificity (τ) for the two probe sets is 0.351 and 0.334, respectively. For the similarity between the two profiles generated by the two probe sets, *ECI* = 0.250 and *r* = 0.849. **(b)** Expression profiles of human *NEU1* gene (probe set: 208926_at) and its mouse ortholog (probe set: gnf1m23979_at). The *ECI* value between the profiles of human-mouse *NEU1* orthologs is 0.961, while *r* is 0.288.

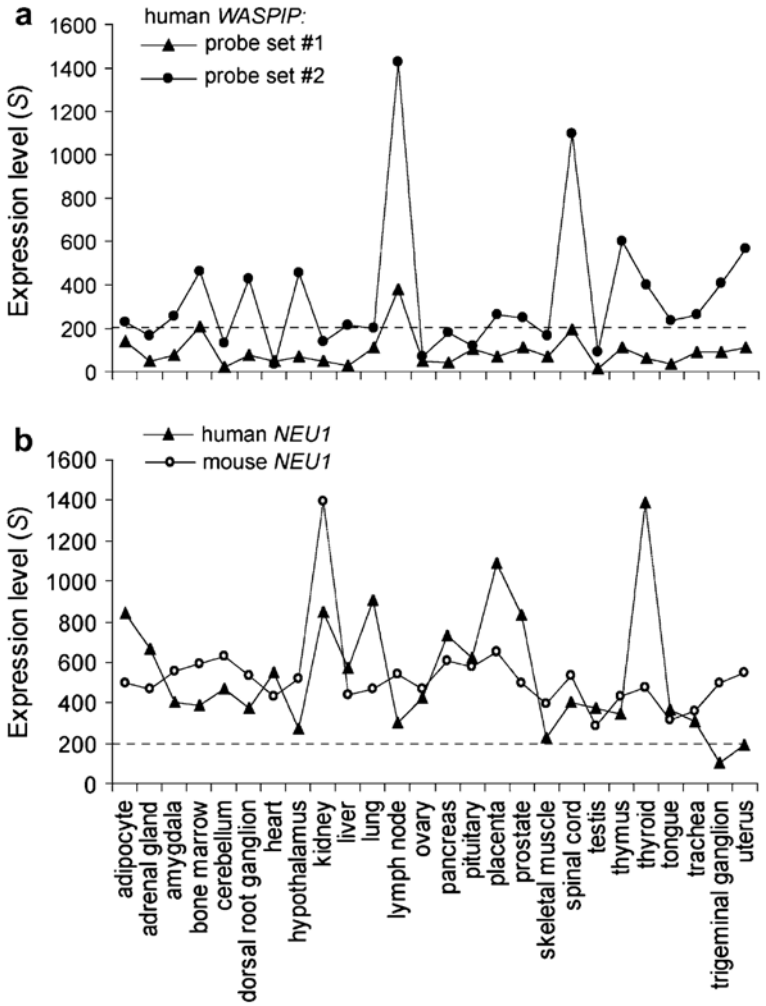


Figure 2.5 The comparison of parameters measuring gene expression conservation and expression breadth. The correlation between expression breadth (*B*) and expression conservation index (*ECI*) is due to the intrinsic dependence between the two parameters. **(a)** *B* and *ECI* are positively correlated in both real orthologs and randomly paired human and mouse genes. Following the procedure that Yang, Su and Li (2005) used to generate their Fig. 3, we calculated *B* from the 47 human tissues that are not studied in mouse. **(b)** Tissue-specificity (τ) and expression-profile similarity (*r*) are positively correlated in real orthologs, but not in randomly paired human and mouse genes.

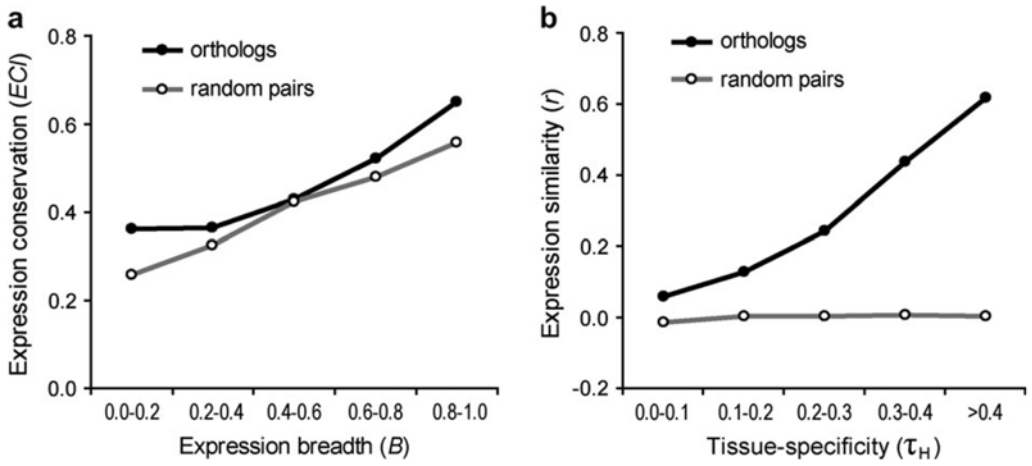
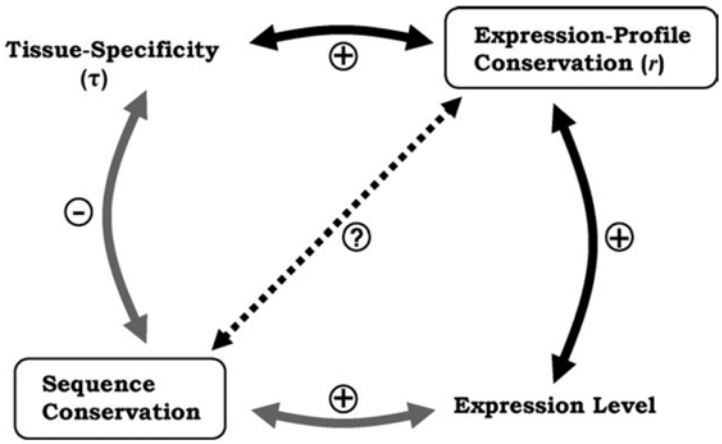


Figure 2.6 A summary of the correlations discussed in this chapter. The "+" symbol denotes a positive correlation; while the "-" symbol denotes a negative correlation. The correlations found in previous studies and confirmed in the present work are presented as grey arrows, while those newly found in this study are presented as black arrows. The relationship between the evolutionary conservation of coding-sequences and that of expression-profiles is unclear.



2.6 LITERATURE CITED

Carroll, S. B. 2005. Evolution at two levels: on genes and form. PLoS Biol **3**:e245.

Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc Natl Acad Sci U S A **97**:12369-12374.

Denver, D. R., K. Morris, J. T. Streelman, S. K. Kim, M. Lynch, and W. K. Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet **37**:544-548.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A Single Determinant Dominates the Rate of Yeast Protein Evolution. Mol Biol Evol **23**:327-337.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102**:14338-14343.

Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17**:68-74.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. Science **296**:340-343.

Gu, X., Z. Zhang, and W. Huang. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci U S A **102**:707-712.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet **18**:609-613.

Gu, Z., S. A. Rifkin, K. P. White, and W. H. Li. 2004. Duplicate genes increase gene expression diversity within and between species. Nat Genet **36**:577-579.

Hastings, K. E. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J Mol Evol **42**:631-640.

He, X., and J. Zhang. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169**:1157-1164.

Hill, A. A., E. L. Brown, M. Z. Whitley, G. Tucker-Kellogg, C. P. Hunter, and D. K. Slonim. 2001. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. Genome Biol **2**:research0055.

Hubbell, E., W. M. Liu, and R. Mei. 2002. Robust estimators for expression analysis. Bioinformatics **18**:1585-1592.

Huminiecki, L., and K. H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res **14**:1870-1879.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**:249-264.

Jordan, I. K., L. Marino-Ramirez, and E. V. Koonin. 2005. Evolutionary significance of gene expression divergence. Gene **345**:119-126.

Jordan, I. K., L. Marino-Ramirez, Y. I. Wolf, and E. V. Koonin. 2004. Conservation and coevolution in the scale-free human gene coexpression network. Mol Biol Evol **21**:2058-2070.

Kasprzyk, A., D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res **14**:160-169.

Khaitovich, P., I. Hellmann, W. Enard, K. Nowick, M. Leinweber, H. Franz, G. Weiss, M. Lachmann, and S. Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Khaitovich, P., G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol **2**:682-689.

King, M. C., and A. C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet **31**:180-183.

Li, W.-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Liao, B.-Y., and J. Zhang. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol **23:**530-540.

Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res **13**:1638-1645.

Murphy, W. J., P. A. Pevzner, and S. J. O'Brien. 2004. Mammalian phylogenomics comes of age. Trends Genet **20**:631-639.

Nei, M., and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Oleksiak, M. F., G. A. Churchill, and D. L. Crawford. 2002. Variation in gene expression within and among natural populations. Nat Genet **32**:261-266.

Pal, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics **158**:927-931.

Ponger, L., L. Duret, and D. Mouchiroud. 2001. Determinants of CpG islands: expression in early embryo and isochore structure. Genome Res **11**:1854-1860.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science **300**:1742-1745.

Rifkin, S. A., J. Kim, and K. P. White. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet **33**:138-144.

Rifkin, S. A., D. Houle, J. Kim, and K. P. White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**:220-223.

Rocha, E. P., and A. Danchin. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol **21**:108-116.

Schug, J., W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, Jr. 2005. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol **6**:R33.

Springer, M. S., W. J. Murphy, E. Eizirik, and S. J. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A **100**:1056-1061.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci U S A **99**:4465-4470.

Su, A. I., T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Subramanian, S., and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**:373-381.

Vinogradov, A. E. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet **20**:248-253.

Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102**:5483-5488.

Winter, E. E., L. Goodstadt, and C. P. Ponting. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res **14**:54-61.

Wu, Z., R. A. Irizarry, R. Gentleman, F. Martinez Murillo, and F. Spencer. 2004. A model based background adjustment for oligonucleotide expression arrays. J Am Stat Assoc **99:**909-917.

Yanai, I., H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21**:650-659.

Yanai, I., D. Graur, and R. Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15-24.

Yang, J., A. I. Su, and W.-H. Li. 2005. Gene Expression Evolves Faster in Narrowly than in Broadly Expressed Mammalian Genes. Mol Biol Evol **22**:2113-2118.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555-556.

Zhang, J., and X. He. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol **22**:1147-1155.

Zhang, L., and W. H. Li. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol **21**:236-239.

CHAPTER 3

IMPACT OF GENE EXPRESSION AND OTHER PROPERTIES OF GENES ON THE RATE OF MAMMALIAN PROTEIN EVOLUTION

3.1 ABSTRACT

Understanding the determinants of the rate of protein sequence evolution is of fundamental importance in evolutionary biology. Many recent studies have focused on the yeast because of the availability of many genome-wide expressional and functional data. Yeast studies revealed a predominant role of gene expression level and a minor role of gene essentiality in determining the rate of protein sequence evolution. Whether these rules apply to complex organisms such as mammals is unclear. Here we assemble a list of 1,642 essential and 1,341 nonessential mouse genes based on targeted gene deletion experiments and report a significant impact of gene essentiality on the rate of mammalian protein evolution. Gene expression level has virtually no effect, although tissue-specificity in expression pattern has a strong influence. Unexpectedly, gene compactness, measured by average intron size and UTR (untranslated region) length, has influence as great as gene essentiality. Hence, the relative importance of the various factors in determining the rate of mammalian protein evolution is gene compactness ≈ gene essentiality > tissue-specificity > expression level. Our results suggest a considerable variation in rate determinants between unicellular organisms such as the yeast and multicellular organisms such as mammals.

3.2 INTRODUCTION

What determines the rate of protein sequence evolution is a fundamental question in molecular evolution. It is well known that the evolutionary rates of different proteins in a genome vary by several orders of magnitude (Dayhoff 1972; Li 1997). This variation is typically explained by differences in the mutation rate and selection intensity among genes (Kimura 1983; Li 1997). However, the biological factors underlying such differences had not been examined with sufficiently large data until a few years ago when genome sequences and functional genomic data became available. Factors that have been shown to influence the protein evolutionary rate include gene essentiality (Hirsh and Fraser 2001; Jordan et al. 2002; Wall et al. 2005; Zhang and He 2005), gene expression level (Pal, Papp, and Hurst 2001b; Akashi 2003; Rocha and Danchin 2004; Subramanian and Kumar 2004; Drummond, Raval, and Wilke 2006), tissue-specificity (Hastings 1996; Duret and Mouchiroud 2000; Subramanian and Kumar 2004; Winter, Goodstadt, and Ponting 2004; Zhang and Li 2004), presence of a duplicate copy (Nembaware et al. 2002; Castillo-Davis and Hartl 2003; Yang, Gu, and Li 2003), properties in the protein-interaction network (Fraser et al. 2002; Fraser 2005; Hahn and Kern 2005; Makino and Gojobori 2006), local recombination rate (Pal, Papp, and Hurst 2001b), and pleiotropy (He and Zhang 2006), although some of these factors are interrelated. In the past few years, many studies have focused on unicellular organisms, particularly the yeast *Saccharomyces cerevisiae*, due to the early availability of a large amount of functional genomic data for this model organism. With the advancement of mammalian genomics, it becomes possible to conduct genome-wide analysis of several biological factors that potentially influence the rate of mammalian protein evolution and to

compare the relative importance of these factors in yeasts and mammals, respective representatives of unicellular and multicellular eukaryotes.

Among the potential rate determinants, gene essentiality is perhaps the most studied and debated factor. Essential genes refer to those that cause lethality or infertility when deleted. Based on the neutral theory of molecular evolution (Kimura and Ohta 1974), it was predicted that essential genes are subject to stronger selective constraints and therefore evolve more slowly than nonessential genes (Wilson, Carlson, and White 1977). However, Hurst and Smith (1999) failed to verify this prediction when they compared 67 essential genes and 108 nonessential genes of the mouse. Although subsequent analysis of bacterial and yeast genes found gene essentiality to be an important rate determinant (Hirsh and Fraser 2001; Jordan et al. 2002), these results were suggested to arise from a confounding factor of the gene expression level (Pal, Papp, and Hurst 2003; Rocha and Danchin 2004). More recent analyses, however, showed that gene essentiality has a small, yet statistically significant, impact on the evolutionary rate of yeast proteins even when the gene expression level is controlled for (Zhang and He 2005; Wall et al. 2005). Nonetheless, despite the availability of many mouse strains produced in targeted gene deletion experiments, whether gene essentiality influences mammalian protein evolution remains unsolved due to the lack of a comprehensive list of essential and nonessential genes.

The importance of gene expression level in determining the protein evolutionary rate in yeasts and bacteria is well established (Pal, Papp, and Hurst 2001a; Rocha and Danchin 2004; Zhang and He 2005; Drummond, Raval, and Wilke 2006), although the molecular evolutionary mechanisms are unclear and debated (Akashi 2003; Drummond et al. 2005). Unlike unicellular organisms, mammalian cells are highly differentiated and different types

of cells turn on different sets of genes to maintain their identities and functions. Hence, both the expression level and tissue-specificity of expression may be important in determining the rate of mammalian gene evolution. In fact, previous studies of mammalian genes showed higher evolutionary rates among lowly expressed genes than highly expressed genes (Subramanian and Kumar 2004) and higher rates among tissue-specific genes than housekeeping genes (Duret and Mouchiroud 2000; Winter, Goodstadt, and Ponting 2004; Zhang and Li 2004). However, because housekeeping genes tend to be highly expressed (Vinogradov 2004; Liao and Zhang 2006a), it is unknown whether expression level and tissue specificity have independent influences on the evolutionary rate.

A previous study of 363 mouse and rat genes showed a significant, but weak, negative correlation between protein length and the rate of protein sequence evolution (Zhang 2001). An opposite pattern, however, was found in the fruitfly (Lemos et al. 2005). Recent studies also showed that highly expressed genes tend to code for short proteins and have short introns (Castillo-Davis et al. 2002). Because highly expressed genes tend to have low rates of protein evolution, one would expect a positive correlation between protein (or intron) length and the rate of protein evolution. It is interesting to test this prediction.

In the present study, we first compile a list of 2,982 mouse genes with essentiality information derived from targeted gene deletion data. We then study the influences of gene essentiality, gene expression level, tissue-specificity, and gene compactness (in terms of protein length, average intron length, and UTR length) on the rate of mammalian protein evolution. We conduct a series of partial correlation analyses to disentangle the contributions of various factors and compare our results with findings from the yeast. Our results reveal a great variation in rate determinants between unicellular and multicellular organisms.

3.3 MATERIALS AND METHODS

3.3.1 Mouse essential and nonessential genes

Mouse genes subject to targeted deletion experiments were downloaded from Mouse Genome Informatics (MGI) (MGI 3.51; http://www.informatics.jax.org/). Only those genes having one corresponding Ensembl gene name were kept for subsequent analysis. These genes were classified into essential and nonessential genes based on their targeted deletion phenotypic codes (MP numbers) provided by MGD. By definition, essential genes are those with the knockout phenotype of lethality or sterility. That is, those entries possessing embryonic lethality (MP: 0002080), prenatal lethality (MP: 0002081), post-natal lethality (MP: 0002082), premature death or induced morbidity (MP: 0002083), abnormal reproductive system morphology (MP: 0002160) or abnormal reproductive system physiology (MP: 0001919) were grouped as essential genes. All other genes associated with a phenotypic classification term, including those entries with a normal phenotype, were grouped as nonessential genes.

3.3.2 Gene orthology and evolutionary rate

The homology information of mouse and rat genes was obtained from Ensembl EnsMart (http://www.ensembl.org/Multi/martview). There were several annotated homology relationships between mouse and rat genes by Ensembl. We only considered those pairs of genes annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BLASTZ searches) to be orthologous. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous

substitutions per nonsynonymous site (d_N) between mouse and rat orthologs were estimated by the maximum likelihood method of Yang (1997) and retrieved from Ensembl EnsMart.

3.3.3 Structural and functional annotations of mouse genes

The structural and functional annotations of mouse genes were obtained from Ensembl version 38. Chromosomal positions, CDS (coding sequence) lengths, intron numbers, intron lengths, and 5'- and 3'-UTR (untranslated region) lengths of mouse genes were retrieved from Ensembl EnsMart (http://www.ensembl.org/Multi/martview) (Kasprzyk et al. 2004). For alternatively spliced genes, we chose structural information of the splice form with the longest coding sequence. Genes having immune-related functions were identified from the Gene Ontology description (http://www.geneontology.org/) contained in Ensembl database. It should be noted that not all mouse genes in the preliminary dataset have rat orthologs. After removing mouse genes without UBRH rat orthologs, 1,642 essential and 1,341 nonessential mouse genes were kept for subsequent analysis.

The gene structure annotation of the yeast *S. cerevisiae* was also obtained from Ensembl EnsMart. Nucleotide substitution rates between *S. cerevisiae* and *S. bayanus* orthologous genes were obtained from Zhang and He (2005).

3.3.4 Analysis of gene expression pattern

The spatial expression information of mouse genes was obtained from the Gene Atlas V2 dataset (http://symatlas.gnf.org/SymAtlas/). This dataset was generated by hybridization of RNAs from 61 mouse tissues onto Affymetrix microarray chips (GNF1M) (Su et al. 2004). To assign expression data from probe sets to corresponding Ensembl mouse genes, we

aligned probe sequences of each probe set to the Ensembl cDNA sequences (Mus_musculus.NCBIM33.feb.cdna.fa; http://www.ensembl.org/info/data/download.html) using BLASTn (http://www.ncbi.nlm.nih.gov/blast/). Only those probe sets in which all matching probes perfectly matched to the same Ensembl gene were considered to be valid. The expression level detected by each probe set was obtained as the signal intensity (*S*) computed from MAS 5.0 algorithm (MAS5) (Hubbell, Liu, and Mei 2002). The *S* values were averaged among replicates.

In the present study, we measured two properties of the mouse gene expression pattern: expression level (*ExpLev*) and tissue-specificity (τ). *ExpLev* is defined as the average signal intensity (*S*) of a mouse gene across 61 examined tissues. The tissue-specificity of a gene is defined as the heterogeneity of its expression level across all the tissues and is estimated by $\tau = \frac{\sum_{j=1}^{n}(1-\left[\frac{\log_2 S(j)}{\log_2 S_{\max}}\right])}{n-1}$, where n=61 is the number of mouse tissues examined here and $S_{\max}$ is the highest expression signal of the gene across all tissues (Yanai et al. 2005). To minimize the influence of noise from low intensities, we arbitrarily let $S(j)$ be 100 if it is lower than 100 (Liao and Zhang 2006a). The τ value ranges from 0 to 1, with higher values indicating greater variations in expressional level across tissues and thus higher tissue specificity. The advantage of using τ rather than expression breadth, which requires an arbitrary cutoff to determine whether a gene is expressed in a given tissue, has been extensively discussed (Liao and Zhang 2006a). Some mouse genes are represented by more than one probe set on the microarray. Because it was not possible to tell which probe set provides the best expression measure of a target gene (Liao and Zhang 2006b), we computed *ExpLev* and τ by averaging the values derived from the different probe sets of the

same gene. The final dataset used in partial correlation analyses contained 2,214 mouse genes with knockout phenotypes, orthologous rat genes, and structural and expression data. Among them, 1,255 were essential and 959 were nonessential.

3.4 RESULTS

3.4.1 Nonessential proteins evolve faster than essential proteins

We compiled a list of essential and nonessential genes using mouse targeted gene deletion data. Among them, 1,642 essential and 1,341 nonessential genes have orthologous genes in the rat. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) were estimated for these genes using mouse and rat orthologs. We found a significant difference between essential and nonessential genes in d_N ($P < 10^{-28}$, Mann-Whitney U test; Figure 3.1a). On average, d_N is 40% greater for nonessential genes than essential genes. We noticed that X-linked genes and immune-system genes are slightly overrepresented in the nonessential group (3.3% and 7.0%), compared to the essential group (2.3% and 3.1%). Because X-linked mammalian genes may behave differently from autosomal genes due to differences in gene content, mutation rate, and selection intensity (Wang et al. 2001; Malcom, Wyckoff, and Lahn 2003; Lu and Wu 2005) and immune-system genes tend to be under diversifying positive selection (Hughes and Nei 1988; Hughes 1999), we repeated the above analysis by removing X-linked genes and immune-related genes. Our results, however, remain unchanged (Figure 3.1a). Although d_S is also significantly higher for nonessential genes than essential genes, the difference in mean d_S between the two groups is small (~3%) (Figure

3.1b). The average d_N/d_S ratio of nonessential genes is 33-42% greater than that of essential genes, depending on whether X-linked genes and immune-system genes are considered or not (Figure 3.1c). Thus, the correlation between gene essentially and d_N or d_N/d_S is significantly negative (Table 3.1). These results indicate that gene essentiality affects the rate of mammalian protein evolution by influencing the selective constraint on the proteins.

3.4.2 Effects of gene expression level and tissue specificity on the rate of protein evolution

Two gene expression properties, expression level (Pal, Papp, and Hurst 2001a; Rocha and Danchin 2004; Subramanian and Kumar 2004; Zhang and He 2005; Drummond, Raval, and Wilke 2006) and tissue-specificity (Hastings 1996; Duret and Mouchiroud 2000; Subramanian and Kumar 2004; Zhang and Li 2004), have been shown to affect the rate of protein sequence evolution to various degrees in different species. Specifically, highly expressed genes and non-tissue-specific genes tend to evolve slowly. Analysis based on our dataset confirms these findings (Table 3.1 and Figure 3.2). Interestingly, although gene expression level is the most important rate determinant in bacteria (Rocha and Danchin 2004) and yeast (Drummond, Raval, and Wilke 2006), the correlation between gene expression level (*ExpLev*) and d_N is weak (Spearman's $\rho = -0.05$) and only marginally significant ($P = 0.01$) in mammals. Similar results are obtained when essential and nonessential genes are analyzed separately. By contrast, the correlation between tissue-specificity (τ) and d_N is much stronger ($\rho = 0.166$, $P < 10^{-16}$). We noticed that tissue-specific genes not only have greater d_N/d_S but also greater d_S values (Figure 3.2), implying that faster protein evolution of tissue-specific genes may have resulted from both higher mutation rate and lower purifying

selection. Since average d_S does not exhibit the same magnitude of increase as average d_N while τ becomes larger (~17% increase versus ~90% increase), mutation rate bias is unlikely to be the main cause for high d_N of tissue-specific genes. Our result is consistent with that of Zhang and Li (2003).

Because the expression level and tissue-specificity may be correlated, we measured the partial correlation between *ExpLev* and d_N by controlling for τ. Although the partial correlation becomes stronger and more significant ($\rho = -0.061$, $P < 10^{-2}$), it is still not comparable to the partial correlation between τ and d_N when *ExpLev* is controlled for ($\rho = 0.174$, $P < 10^{-18}$). These results suggest that tissue-specificity is much more important than average expression level in determining the rate of mammalian protein sequence evolution.

3.4.3 Compact genes have high rates of evolution

Although a significant positive correlation between the CDS length and d_N was observed in fruitfly (Lemos et al. 2005) and a significant negative correlation was observed in a set of 363 mouse and rat genes (Zhang 2001), no significant correlation is found in our data (Table 3.1). Surprisingly, we found a negative correlation between UTR length and d_N (or d_N/d_S) (Figure 3.3 and Table 3.1). For example, the mean d_N of genes with a total UTR length of <300 nucleotides is about twice that of genes with a total UTR length of >2400 nucleotides (Figure 3.3a). Similarly, we found a negative correlation between average intron size in a gene (but not intron number) and d_N (or d_N/d_S) of the gene (Figure 3.4 and Table 3.1). The mean d_N of genes with an average intron size of <1000 nucleotides is over 5 times that of genes with an average intron size of >8000 nucleotides (Figure 3.4a). The

correlations between gene compactness and d_N are of comparable or even higher magnitudes than that between tissue-specificity (τ) and d_N (Table 1).

In the above analysis, we used the longest splice form for those genes that have alternative splicing. We repeated the above analysis by using the shorted splice form or removing genes with alternative splicing. The results are essentially the same (Tables A.1 and A.2). There are also many overlapping (including nested) genes in the mouse genome (Veeramachaneni et al. 2004). Removing these genes does not affect our result (Table A.3).

3.4.4 Relative impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate

The above examined factors are not completely independent in determining the rate of protein sequence evolution. For instance, genes with high expression levels tend to have small introns ($\rho = -0.079$, $P < 10^{-4}$). In order to separate the contributions of multiple factors, we applied partial correlation analyses. Although a recent study suggested that principle component analysis is superior to partial correlation analysis for noisy data (Drummond, Raval, and Wilke 2006), subsequent analytical and empirical analyses do not support this view (S. Yi, personal communication). In our partial correlation analysis, we focused on the correlation between the evolution rate and one of the three factors (i.e., gene essentiality, expression pattern, and gene compactness), by controlling the other two factors. All factors having very significant effects (P<0.01) on the evolutionary rate in Table 3.1 show significant and independent effects on d_N and d_N/d_S, with the exception of *ExpLev* (Table 3.2). After controlling for gene essentiality, the negative correlation between *ExpLev* and d_N and that between *ExpLev* and d_N/d_S become only marginally significant ($P = 0.041$ and 0.026,

respectively), suggesting that the weak negative correlation between gene expression level and protein evolutionary rate in Table 3.1 may be due to the fact that essential genes tend to have both high *ExpLev* and low d_N. Our result thus suggests that the effect of expression level itself on the evolutionary rate of mammalian proteins is negligible. We notice that genes with high expression levels tend to have high d_S but low d_N/d_S (Table 3.2). Hence, the relatively weak correlation between *ExpLev* and d_N may be due to the opposite effects of high mutation rates and strong purifying selection at highly expressed genes. Comparing to properties of gene expression (expression level and tissue specificity), gene essentiality and compactness seem to have larger impacts on the rate of mammalian protein evolution (Table 3.1 and Table 3.2). Based on the partial correlation analysis (Table 2), we conclude that the relative importance of the factors in determining the rate of mammalian protein evolution is gene compactness ≈ gene essentiality > tissue-specificity > gene expression level.

3.5 DISCUSSIONS

In this work, we used statistical analysis to study the determinants of the rate of mammalian protein sequence evolution. Because there are potentially many rate determinants and because some measures of these determinants (e.g., gene expression level and tissue specificity) have large estimation errors (Wall et al. 2005; Liao and Zhang 2006b), it is not unexpected that the observed correlation coefficients are not very high. We thus evaluate the impact of each factor by considering both the statistical significance in correlation analysis and the magnitude of the correlation. We also compare the impacts of different factors for a given species and the impacts of the same factor across species.

Based on an analysis of 175 mouse genes, Hurst and Smith (1999) found no significant correlation between gene essentiality and d_N/d_S. Zhang and He (2005) suggested that this negative result was likely due to an insufficient sample size. Indeed, when 2,983 mouse genes are analyzed here, essential genes showed significantly lower d_N/d_S than nonessential genes. This difference remains highly significant even when we remove immune-system genes and X-linked genes. Furthermore, the correlation between gene essentiality and d_N (or d_N/d_S) is still significant after controlling for gene expression level, tissue specificity, UTR length, and intron length. We conclude that gene essentiality is an independent determinant of the rate of mammalian protein evolution. It is interesting to note that in yeasts, the average d_N of nonessential genes is ~40% higher than that of essential genes (Zhang and He 2005), a number slightly greater than that observed for mammalian genes (30%). The rank correlation coefficient between gene essentiality and d_N is ~0.2 in yeast, also slightly greater than that in mammals (0.14). After controlling for gene expression level, the correlation coefficient becomes 0.10-0.15 in yeast and 0.17 in mammals. Note that the yeast gene knockout data used by Zhang and He (2005) contained >90% of yeast genes, while the mouse gene knockout data used here contained only 15% of mouse genes. Since targeted gene deletion in mouse requires great efforts, it is possible that researchers tend to study and report functionally important mouse genes that have human orthologs, thus reducing the variation in essentiality among the genes included in our dataset. This reduction could potentially decrease the correlation coefficient between gene essentiality and d_N. But, at any rate, gene essentiality and d_N are significantly correlated in mammals. Thus, in all organisms so far examined (bacteria, yeasts, nematodes, and mammals), nonessential genes tend to evolve faster than essential genes. It is thus

appropriate to conclude that the fundamental prediction of the neutral theory, that less important genes evolve faster than important genes, is universally supported by empirical data at the genomic level. However, it should be pointed out that the correlation between gene essentiality and d_N, although statistically significant, is small in magnitude. This weak correlation contrasts the strong belief of many biologists that functionally important DNA sequences evolve slowly, which is the basis of many successful bioinformatic methods such as BLAST (Altschul et al. 1990) and phylogenetic footprinting (Gumucio et al. 1993). It is possible that the knockout phenotype observed in the lab only roughly reflects the amount of fitness reduction in the wild, which is expected to be a better rate determinant.

A previous study showed that human morbid genes (those known to cause diseases when mutated) evolve more slowly than non-morbid genes (Kondrashov, Ogurtsov, and Kondrashov 2004). Their analysis is not equivalent to a comparison between essential and nonessential genes, because non-morbid genes can have unidentifiable embryonic lethal phenotype or infertility phenotype when mutated. In other words, non-morbid genes include both essential and nonessential genes and thus there is no clear prediction as whether non-morbid genes should evolve more rapidly or more slowly than morbid genes. In fact, Smith and Eyre-Walker (2003) also analyzed morbid and non-morbid genes, but obtained an opposite result.

We found that the rate of mammalian protein evolution is not, or is only weakly, correlated with the gene expression level, when gene essentiality is controlled for. In the future, it would be important to verify this finding for the entire genome as more gene knockout data become available. If our finding is generally true for mammals, it contrasts that from the yeast, where the expression level explains about a quarter (ρ^2=~0.25) of the

variation in d_N (Zhang and He 2005). The reduction of the correlation in mammals may be due to smaller population sizes in mammals than in yeasts, because the expression level becomes a weaker selective force as the population size reduces (Ohta 1992). However, although the correlations between various rate determinants and protein evolutionary rate in mammals may be reduced due to smaller population sizes, the relative importance of these rate determinants should remain unchanged. Why are the influences of gene expression level on d_N drastically different between yeast and mammals? To address this question, one has to understand why the gene expression level affects d_N in yeast. However, no widely accepted explanation exists at this time. The recently proposed translational robustness hypothesis (Drummond et al. 2005) suggests that highly expressed proteins are prone to forming misfolded protein aggregates that could be toxic or pathogenic to the organism (Ellis and Pinheiro 2002). Thus, their coding regions are under intense selective pressure to maintain certain sequences that avoid misfolding in the presence of translational errors (Drummond et al. 2005). If this hypothesis is correct, our observation of no impact of expression level on d_N in mammals may be due to a lowered probability of protein aggregation in mammalian cells. It is known that a misfolded protein may aggregate, particularly when it is in high concentration (Minton 2000). The cell volume of the mouse sperm (61-70μm^3) (Brotherton 1975), the smallest mouse cell, is similar to that of a haploid yeast cell (~70μm^3) (Sherman 1991). Generally speaking, other types of mammalian cells are much larger than the sperm cell (and the yeast cell). If the protein concentration (per gene) in a cell is generally lower in mammals than in yeast, the pressure of avoiding aggregation would also be lower in mammalian cells, making expression level a negligible factor in determining d_N. Nonetheless, this explanation is built on two assumptions, the translational robustness

hypothesis and a lower protein concentration per gene in mammalian cells than in yeast cells, both of which require further scrutiny. An alternative explanation is that the gene expression level of a unicellular organism and the average gene expression level across tissues of a multicellular organism are two different things and are not comparable. Interestingly, when using the gene expression level estimated from the mouse ESTs at an embryonic stage, Subramanian and Kumar (2004) found a significant impact of gene expression level on the rate of protein evolution. Because many genes are not expressed at the embryonic stage, the biological meaning of their observation is not immediately clear. It remains to be seen whether the correlation between gene expression level and protein evolutionary rate exists only among genes having similar functions or expression patterns (as in Subramanian and Kumar's study), but not among genes with diverse properties. Alternatively, the microarray gene expression data used in the present study may be too noisy to accurately reflect mRNA abundance compared to the EST data used by Subramanian and Kumar (2004). But, interestingly, the same microarray data revealed a strong correlation between τ and d_N, suggesting that these data still contain a sufficient amount of expression information. We also examined the correlation between the d_N of a gene and the maximum expression level of the gene across 61 tissues surveyed. Unexpectedly, a weak positive correlation was observed (ρ=0.075, P=1.4×10^{-4}). It is unclear what caused this positive correlation.

A surprising finding of the present study is that compact genes (with short UTRs and introns) tend to evolve fast (Figure 3.3 and 3.4). Although the above finding was based on genes with knockout data, essentially the same result was obtained when the entire genome is analyzed (Table A.4). Previous studies showed that highly expressed genes have short introns (Castillo-Davis et al. 2002) and evolve slowly (Subramanian and Kumar 2004). Thus,

one expects that genes with short introns evolve slowly. But, our observation is opposite. The reason for this unexpected observation is not entirely clear. Of course, in our analysis, gene expression level and d_N are virtually uncorrelated, and thus the prediction that compact genes evolve slowly is invalid. Nevertheless, the observation that compact genes evolve fast is still surprising. Since UTRs and introns are noncoding regions of a gene and the majority of these sequences are more tolerant than coding regions to insertions and deletions, we consider the length variation of these noncoding sequences as a result of variation of local insertion and deletion rates (Vinogradov 2004). That is, we assume that the insertion/deletion rate ratio varies across genomic regions, making some genes more compact than others. It has been proposed that the presence and length of noncoding regions such as introns and intergenic regions can increase the frequency of recombination between adjacent exons and genes (Comeron and Kreitman 2002). Accordingly, for two genes with the same functional importance, same CDS length, same number of introns, but different intron sizes, purifying selection is expected to be more efficient for the gene with bigger introns than the one with smaller introns, as the former has a higher recombination rate (per gene) than the latter. This difference results in a lower expected d_N for the gene with bigger introns, which is observed in this study. Of course, recombination rate variation provides just one possible explanation of our observation; other possibilities cannot be excluded. Contrary to mammals, only 263 yeast protein-coding genes (~5%) contain intron(s). Thus, it is expected that gene compactness will not be an important factor in determining yeast protein evolution at the genomic level. However, among 86 intron-containing yeast (*S. cerevisiae*) genes that have *S. bayanus* orthologs, the average intron size and d_N are negatively correlated ($\rho = -0.282$, $P < 0.01$), similar to the result obtained from mammalian genes. It would be interesting to

examine whether the influence of gene compactness on protein evolutionary rate is as significant for unicellular eukaryotes with high prevalence of introns (e.g., the green algae *Chlamydomonas reinhardtii*) as in mammals.

In summary, we find that the relative importance of various rate determinants in mammals is gene compactness ≈ gene essentiality > tissue-specificity > gene expression level. This order differs substantively from that in yeasts or bacteria. For example, although the absolute magnitudes of the impact of gene essentiality are similar between the yeast and mammals, the relative impacts appear quite different, because the gene expression level plays a much greater role in yeast than in mammals. It seems that the rules governing the rate of protein evolution need not be the same for all major clades of living organisms. Our results highlight the danger of applying findings from a single species, even based on a genome-wide analysis, to distantly related species, and suggest reexamination of the roles of various rate determinants across a wide range of species, which is becoming feasible with the rapid advance of functional and comparative genomics.

3.6 ACKNOWLEDGMENTS

We thank Wendy Grus, Ondrej Podlaha, Peng Shi, and two anonymous reviewers for valuable comments. This work was supported by research grants from the University of Michigan and the National Institutes of Health to J.Z.

Figure 3.1 Nonessential mouse genes evolve faster than essential genes. Average mouse-rat **(a)** d_N, **(b)** d_S, and **(c)** d_N/d_S values of essential and nonessential genes are shown. *P*-value from the test of the null hypothesis of no difference between essential and nonessential genes is shown above each comparison (Mann-Whitney *U* test). Error bars represent the standard error of the mean. All genes: 1,612 essential and 1,341 nonessential. Non-immune-system genes: 1,538 essential and 1,173 nonessential. Autosomal genes: 1,597 essential and 1,290 nonessential. Autosomal, non-immune-system genes: 1,494 essential and 1,129 nonessential.

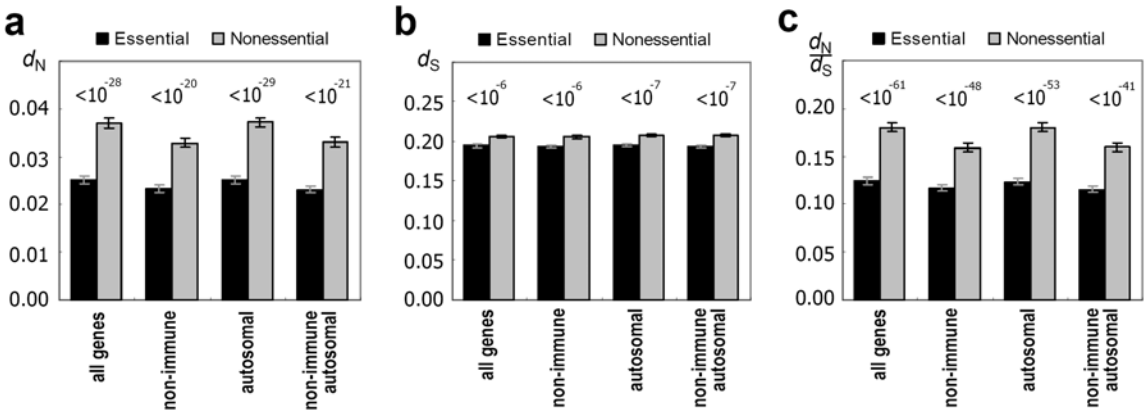


Figure 3.2 Evolutionary rate of mouse genes positively correlates with tissue-specificity (τ). Average mouse-rat **(a)** d_N, **(b)** d_S, and **(c)** d_N/d_S values of each bin are shown. Error bars represent the standard error of the mean.

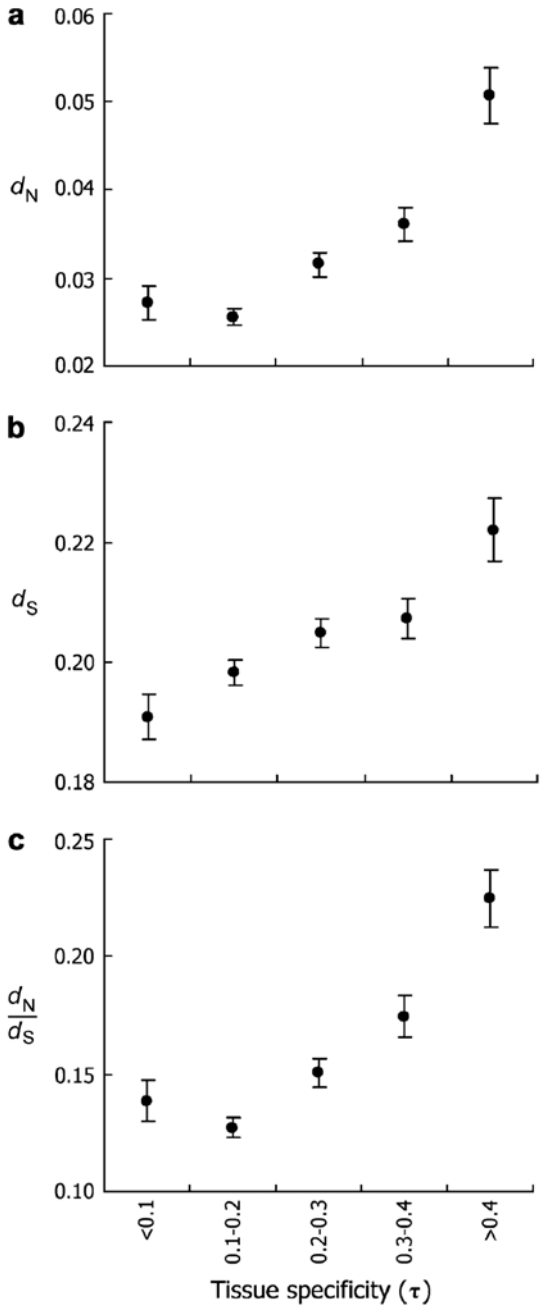


Figure 3.3 Mouse genes with longer UTRs (untranslated regions) tend to have lower d_N and d_N/d_S values. Average mouse-rat **(a)** d_N, **(b)** d_S, and **(c)** d_N/d_S values of each bin are shown. Error bars represent the standard error of the mean.

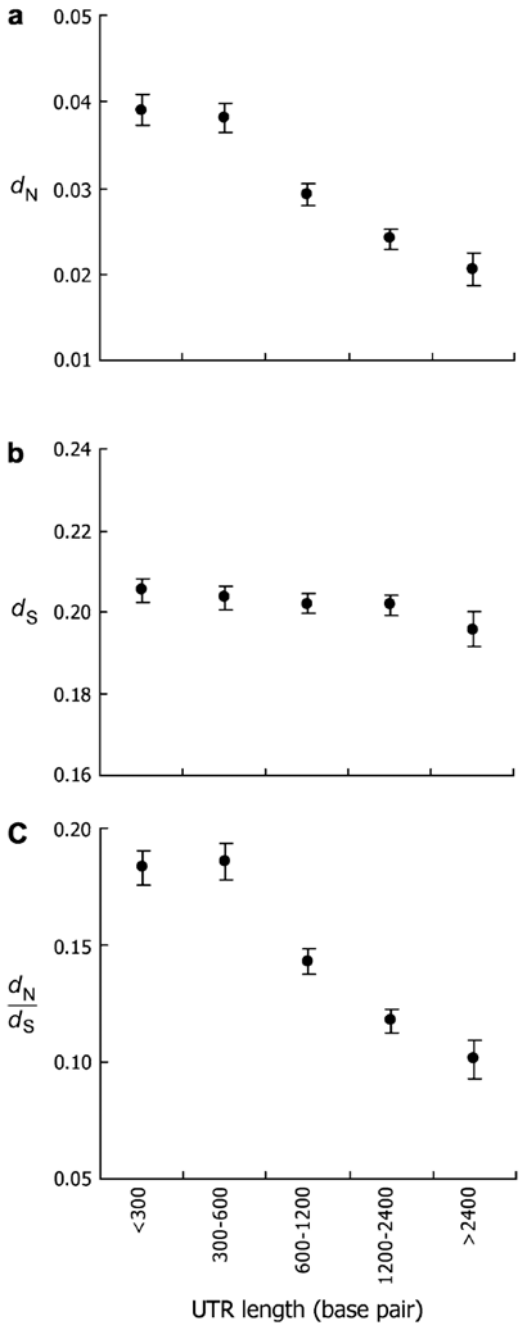


Figure 3.4 Mouse genes with larger average intron size tend to have lower d_N and d_N/d_S values. Average mouse-rat **(a)** d_N, **(b)** d_S, and **(c)** d_N/d_S values of each bin are shown. Error bars represent the standard error of the mean.

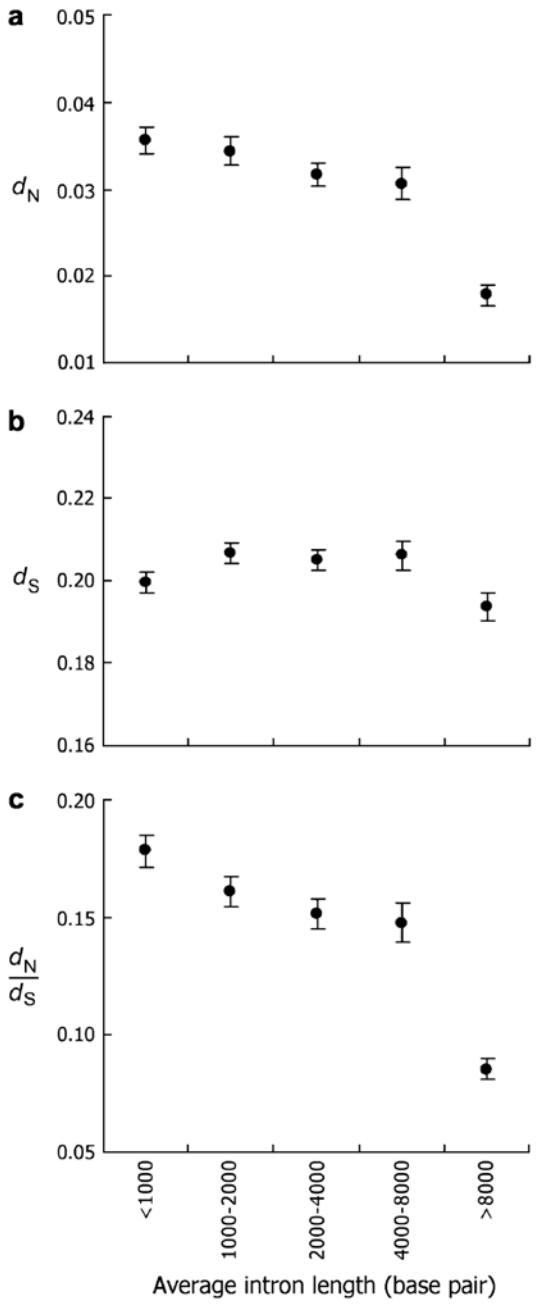


Table 3.1 Spearman's rank correlation coefficient (ρ) between various factors and d_N, d_S or d_N/d_S.

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Essentiality	-0.18982	2.04E-19	-0.10961	2.34E-07	-0.16634	3.35E-15
Expression pattern						
ExpLev	-0.05419	1.08E-02	0.01321	5.35E-01	-0.05691	7.40E-03
τ	0.16550	4.66E-15	0.10156	1.68E-06	0.14827	2.36E-12
Gene structures						
CDS length	0.06003	4.72E-03	0.07360	5.28E-04	-0.03684	8.31E-02
UTR length	-0.20655	8.98E-23	-0.05098	1.65E-02	-0.20025	1.79E-21
5'-UTR length	-0.13144	5.37E-10	-0.04018	5.87E-02	-0.12991	8.50E-10
3'-UTR length	-0.18623	9.76E-19	-0.03655	8.55E-02	-0.18115	8.54E-18
Intron number	0.07470	4.35E-04	0.08631	4.77E-05	0.06064	1.72E-02
Intron length (avg)	-0.11942	1.74E-08	0.01488	4.84E-01	-0.13444	2.14E-10

Note-Essentiality is 1 for essential genes and 0 for nonessential genes. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,214 mouse genes and their rat orthologs.

Table 3.2 Partial rank correlation of various factors and d_N, d_S or d_N/d_S.

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Essentiality						
Essentiality | τ · *ExpLev*	-0.17592	7.56E-17	-0.10304	1.18E-06	-0.15281	4.89E-13
Essentiality | UTR · Intron	-0.18822	4.20E-19	-0.10722	4.26E-07	-0.16484	5.94E-15
Expression pattern						
τ | Essentiality	0.15373	3.52E-13	0.09374	9.96E-06	0.13744	8.35E-11
τ | UTR · Intron	0.13744	8.35E-11	0.09611	5.89E-06	0.12003	1.47E-08
ExpLev | Essentiality	-0.04331	4.16E-02	0.02015	3.43E-01	-0.04736	2.59E-02
ExpLev | UTR·Intron	-0.05249	1.35E-02	0.01707	4.22E-01	-0.05668	7.64E-03
Gene compactness						
UTR length | τ · *ExpLev*	-0.18377	2.87E-18	-0.03664	8.48E-02	-0.17940	1.81E-17
5'-UTR length | τ · *ExpLev*	-0.12785	1.56E-09	-0.03777	7.56E-02	-0.12640	2.40E-09
3'-UTR length | τ · *ExpLev*	-0.16575	4.19E-15	-0.02414	2.56E-01	-0.16229	1.55E-14
UTR length | Essentiality	-0.20318	4.66E-22	-0.04713	2.66E-02	-0.19681	9.04E-21
5'-UTR length | Essentiality	-0.12808	1.47E-09	-0.03710	8.09E-02	-0.12670	2.20E-09
3'-UTR length | Essentiality	-0.18386	2.77E-18	-0.03342	1.16E-01	-0.17863	2.48E-17
Intron length (avg) | τ · *ExpLev*	-0.11321	9.25E-08	0.02248	2.90E-01	-0.12949	9.65E-10
Intron length (avg) | Essentiality	-0.12422	4.51E-09	0.01350	5.26E-01	-0.13860	5.77E-11

Note-Essentiality is 1 for essential genes and 0 for non-essential genes. "UTR" for UTR length and "Intron" for average intron length. The factor before "|" is the factor being examined and those after "|" are the factors being controlled for. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,214 mouse genes and their rat orthologs.

3.7 LITERATURE CITED

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. **215**:403-410.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics **164**:1291-1303.

Brotherton J. 1975. The counting and sizing of spermatozoa from ten animal species using a Coulter counter. Andrologia **7**:169-185.

Castillo-Davis CI, Hartl DL. 2003. Conservation, relocation and duplication in genome evolution. Trends Genet **19**:593-597.

Castillo-Davis, CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet **31**:415-418.

Comeron JM, Kreitman M. 2002. Population, evolutionary and genomic consequences of interference selection. Genetics **161**:389-410.

Dayhoff MO. 1972. Atlas of Protein Sequence and Structure. Natl. Biomed. Res. Found., Washington, DC.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102**:14338-14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327-337.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17**:68-74.

Ellis RJ, Pinheiro TJ. 2002. Medicine: danger--misfolding proteins. Nature **416**:483-484.

Fraser HB. 2005. Modularity and evolutionary constraint on proteins. Nat Genet **37**:351-352.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science **296**:750-752.

Gumucio DL, Shelton DA, Bailey WJ, Slightom JL, Goodman M. 1993. Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. Proc Natl Acad Sci U S A **90**:6018-6022.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol **22**:803-806.

Hastings KE. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J Mol Evol **42**:631-640.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics, in press.

Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. Nature **411**:1046-1049.

Hubbell E, Liu WM, Mei R. 2002. Robust estimators for expression analysis. Bioinformatics **18**:1585-1592.

Hughes AL. 1999. Adaptive evolution of genes and genomes. New York: Oxford University Press.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167-170.

Hurst LD, Smith NG. 1999. Do essential genes evolve slowly? Curr Biol **9**:747-750.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res **12**:962-968.

Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. 2004. EnsMart: a generic system for fast and flexible access to biological data. Genome Res **14**:160-169.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kimura M, Ohta T. 1974. One some principles governing molecular evolution. Proc Natl Acad Sci U S A **71**:2848-2852.

Kondrashov FA, Ogurtsov AY, Kondrashov AS. 2004. Bioinformatical assay of human gene morbidity. Nucleic Acids Res **32**:1731-1737.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol **22**:1345-1354.

Li WH. 1997. Molecular evolution. Sunderland: Sinauer Associates.

Liao BY, Zhang J. 2006a. Low rates of expression-profile divergence in highly-expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol **23**:1119-1128.

Liao BY, Zhang J. 2006b. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol **23**:530-540.

Lu J, Wu CI. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci U S A **102**:4063-4067.

Makino T, Gojobori T. 2006. The evolutionary rate of a protein is influenced by features of the interacting partners. Mol Biol Evol **23:**784-789.

Malcom CM, Wyckoff GJ, Lahn BT. 2003. Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Mol Biol Evol **20**:1633-1641.

Minton AP. 2000. Implications of macromolecular crowding for protein assembly. Curr Opin Struct Biol **10**:34-39.

Nembaware V, Crum K, Kelso J, Seoighe C. 2002. Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res **12**:1370-1376.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst **23**:263–286.

Pal C, Papp B, Hurst LD. 2001a. Highly expressed genes in yeast evolve slowly. Genetics **158**:927-931.

Pal C, Papp B, Hurst LD. 2001b. Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. Mol Biol Evol **18**:2323-2326.

Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol **21**:108-116.

Sherman F. 1991. Getting started with yeast. Methods Enzymol **194**:3-21.

Su AI, Wiltshire T, Batalov S, et al. (13 co-authors). 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**:373-381.

Veeramachaneni V, Makalowski W, Galdzicki M, Sood R, Makalowska I. 2004. Mammalian overlapping genes: the comparative perspective. Genome Res. **14:**280-286.

Vinogradov AE. 2004. Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet **20**:248-253.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102**:5483-5488.

Wang PJ, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat Genet **27**:422-426.

Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. Annu Rev Biochem **46**:573-639.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res **14**:54-61.

Yanai I, Benjamin H, Shmoish M, et al. (12 co-authors). 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21**:650-659.

Yang J, Gu Z, Li WH. 2003. Rate of protein evolution versus fitness effect of gene deletion. Mol Biol Evol **20**:772-774.

Yang Z. 1997. PAML: a program packagefor phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13**:555–556.

Zhang J. 2001. Protein-length distributions for the three domains of life. Trends Genet **16**:107-109.

Zhang J, He, X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol **22**:1147-1155.

Zhang L, Li WH. 2004. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol **21**:236-239.

CHAPTER 4

CO-EXPRESSION OF MAMMALIAN LINKED GENES AND ITS IMPACT ON THE EVOLUTION OF GENOME ARCHETECTURE

4.1 ABSTRACT

Similarity in gene expression pattern between closely linked genes is known in several eukaryotes. Two models have been proposed to explain the presence of such co-expression patterns. The adaptive model assumes that co-expression is advantageous and is established by relocation of initially unlinked but co-expressed genes, whereas the neutral model asserts that co-expression is a type of leaky expression due to similar expressional environments of linked genes, but is neither advantageous nor detrimental. However, these models are incompatible with several empirical observations. Here, we propose that co-expression of linked genes is a form of transcriptional interference that is disadvantageous to the organism. We show that even distantly linked genes that are tens of megabases away exhibit significant co-expression in the human genome. However, the linkage is more likely to be broken during evolution between genes of high co-expression than between those of low co-expression and the breakage of linkage reduces gene co-expression. These results support our hypothesis that co-expression of linked genes in mammalian genomes is generally disadvantageous, implying that many mammalian genes may never reach their optimal expression pattern due to the interference of their genomic environment and that such transcriptional interference may be a force promoting recurrent relocation of genes in the genome.

4.2 INTRODUCTION

Nonrandom distribution of genes in a genome, a widespread phenomenon in prokaryotes (Lawrence 1999), has also been observed in various eukaryotes (reviewed in Hurst, Pal and Lercher 2004). In mammals, linked genes sharing similar expression patterns are often referred to as a gene cluster. For example, clusters of highly expressed genes (Caron et al. 2001), tissue-specific genes (Megy, Audic, and Claverie 2003; Versteeg et al. 2003), broadly expressed genes (Lercher, Urrutia, and Hurst 2002), and co-expressed genes (Fukuoka, Inaoka, and Kohane 2004; Singer et al. 2005; Semon and Duret 2006) have been observed in the human genome. The general phenomenon of co-expression of linked genes has also been reported in other model eukaryotes such as the yeast *Saccharomyces cerevisiae* (Cohen et al. 2000; Kruglyak and Tang 2000; Huynen, Snel, and Bork 2001; Fukuoka, Inaoka, and Kohane 2004; Lercher and Hurst 2006), nematode *Caenorhabditis elegans* (Lercher, Blumenthal, and Hurst 2003; Fukuoka, Inaoka, and Kohane 2004), and fruit fly *Drosophila melanogaster* (Boutanaev et al. 2002; Spellman and Rubin 2002; Bailey et al. 2004; Fukuoka, Inaoka, and Kohane 2004; Kalmykova et al. 2005).

However, it is unclear as to how and why linked genes become co-expressed. The observation that genes involved in the same pathway (Lee and Sonnhammer 2003) or protein complex (Teichmann and Veitia 2004) and genes having similar functions (Cohen et al. 2000) tend to be linked suggests that co-expression of linked genes may be important to gene function (Hurst, Williams, and Pal 2002; Singer et al. 2005). This view, referred to as the adaptive model, assumes that it is beneficial for genes that require co-expression to be brought together via chromosomal rearrangement (Miller et al. 2004; Richards et al. 2005; Singer et al. 2005). The model predicts that once a co-expressed gene cluster is established,

the linkage of the co-expressed genes should be evolutionarily maintained by purifying selection (Hurst, Williams, and Pal 2002; Singer et al. 2005).

Observations of functional similarity of co-expressed linked genes would support the adaptive model. However, when protein function is defined by Gene Ontology (GO), a study of *Drosophila* did not find functional similarity among co-expressed neighboring genes (Spellman and Rubin 2002). In humans, clusters of co-expressed linked genes that belong to the same functional category, as defined by GO, are rare (Fukuoka, Inaoka, and Kohane 2004). Furthermore, although the evolutionary conservation of linkage between co-expressed genes in several yeasts supports the adaptive model (Hurst, Williams, and Pal 2002), considering the recent discovery of long-range co-regulation (~100 kilobases, covering ~30 genes) of linked yeast genes (Lercher and Hurst 2006), the adaptive model implies that the gene order in the yeast genome must be highly organized. However, the high plasticity of yeast gene order revealed from a comparison of 11 species (Fischer et al. 2006) argues against this view. In addition, it is well known that chromatin structures control the expression of nearby genes, regardless of whether these genes are functionally related or not (Hurst, Pal, and Lercher 2004; Sproul, Gilbert, and Bickmore 2005). For instance, the *CD79B* antigen gene, which is located between the human growth hormone cluster and its locus control region on chromosome 17, is expressed in the pituitary, although its function appears B-cell-specific (Cajiao et al. 2004). Thus, it is possible that similar expression of linked genes has no adaptive value.

A recent study on mammalian co-expressed linked genes suggested that co-expressed gene clusters are formed by a neutral evolutionary process (Semon and Duret 2006). That is, expression similarity of linked genes is due to transcriptional interference (Eszterhas et al.

2002) and is not necessarily advantageous. Here, transcriptional interference refers to influence of transcription of one gene on the transcription of another gene, and can be due to shared *cis*-regulatory elements or chromatin structures, among other things. Our *ad hoc* use of transcriptional interference is different from a more narrow definition used elsewhere (Shearwin, Callen, and Egan 2005). The neutral model for the formation of co-expressed gene clusters (Semon and Duret 2006) implies that gene expression patterns are not functionally important and thus can change freely during evolution, which is exactly the neutral model of transcriptome evolution (Khaitovich et al. 2004). Although some early studies had favored this neutral model (Khaitovich et al. 2004; Yanai, Graur, and Ophir 2004), these studies were later shown to have either technical problems or alternative interpretations (Liao and Zhang 2006b). On the contrary, there is increasing evidence that a considerable fraction of genes in a genome are evolutionarily conserved in expression (Nuzhdin et al. 2004; Denver et al. 2005; Jordan, Marino-Ramirez, and Koonin 2005; Khaitovich et al. 2005; Rifkin et al. 2005; Liao and Zhang 2006b; Whitehead and Crawford 2006; Xing et al. 2007). Because co-expression of neighboring genes is a widespread phenomenon (Semon and Duret 2006), it is unlikely that such gene clusters can be formed without any influence on fitness.

Hence, neither the adaptive model nor the neutral model can adequately explain the existence of co-expressed gene clusters. Here, we propose that co-expression of linked genes is due to transcriptional interference that is detrimental to the organism. We test our hypothesis in humans, exploiting the availability of a comprehensive spatial gene expression dataset (Su et al. 2004). We examined co-expression patterns of closely and distantly linked genes in humans and counted evolutionary losses of gene linkage using multiple mammalian

genomes. Lower evolutionary conservation of linkage is found for pairs of genes with high co-expression than those with low co-expression, consistent with the predictions of our hypothesis. Based on these findings, we propose a model of the origin and evolutionary dynamics of co-expression of linked genes.

4.3 MATERIALS AND METHODS

4.3.1 Genome data and annotations

The human genome assembly used in the present study is NCBI version 35, in which the position and orthology annotation (to mouse, rat, and dog) of 34,404 known or predicted genes can be found in Ensembl Archive release v37 (http://feb2006.archive.ensembl.org/). Genome annotations were retrieved through BioMart (http://www.biomart.org/). There were several annotated homology relationships between human and other mammalian genes by Ensembl. We only considered homologous gene pairs annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BLASTz searches) to be orthologous. By this definition, 10,500 human autosomal genes were found to have unambiguous orthologs in mouse (NCBI v34), rat (RGSC 3.4) and dog (CanFam 1.0) genomes.

4.3.2 Analysis of the microarray data

We obtained the expression information of human genes and mouse genes from the Gene Atlas V2 dataset (http://symatlas.gnf.org/SymAtlas/) (Su et al. 2004). This dataset comprises oligonucleotide microarray data in 73 human and 61 mouse normal tissues. To

assign the expression data from probe sets to corresponding Ensembl genes, probe sequences of each probe set were aligned to the Ensembl cDNA sequences (human: Homo_sapiens.NCBI35.feb.cdna.fa; mouse: Mus_musculus.NCBIM33.feb.cdna.fa; http://www.ensembl.org/info/data/download.html) using BLASTn (http://www.ncbi.nlm.nih.gov/blast/). Only those probe sets in which all perfect-match (PM) probes perfectly matched to the same Ensembl gene were considered to be valid. The expression level detected by each probe set was obtained as the signal intensity (*S*) computed from MAS 5.0 algorithm (MAS5) (Hubbell, Liu, and Mei 2002). The *S* values were averaged among replicates. It should be noted that some genes are represented by more than one probe set on the microarray. Because it was not possible to tell which probe set provides the best expression measure of a target gene (Liao and Zhang 2006b), we arbitrarily chose the probe set with highest expression level (Jordan, Marino-Ramirez, and Koonin 2005), which was defined by the summation of *S* across all the examined tissues. As a result, 16,457 human and 16,134 mouse Ensembl genes were assigned with microarray gene expression data.

4.3.3 Removal of duplicate genes

Duplicated genes are expected to have similar expression patterns by ancestry, and such genes, if generated by tandem duplication, are often located in physical proximity to one other. The presence of tandem duplicate genes will artificially generate a negative correlation between the expression similarity of two linked genes and the physical distance between them. Furthermore, duplicate genes are subject to the problem of off-target cross-

hybridization in gene expression measurement; removing duplicate genes further eliminates the co-expression pattern artificially generated by cross-hybridization.

We followed the conventional approach (Lercher, Urrutia, and Hurst 2002; Lercher, Blumenthal, and Hurst 2003; Singer et al. 2005) to remove this known artifact: First, to identify proteins belonging to the same gene family, an all-against-all BLASTp search was performed on the entire protein dataset of a genome (for genes having more than one isoform, the longest peptides were used). To be conservative in the analysis, pairs of proteins with BLAST E-values < 0.2 were considered to be members of the same gene family (Lercher, Urrutia, and Hurst 2002). We then generated a duplicate-free dataset by randomly keeping one member of each gene family and removing all other members. Consequently, a subset of 4,857 human autosomal genes that have expression data was retained. By the same approach, a set of 5,384 mouse genes without duplicates was obtained.

Some of our analyses require the use of human genes and their orthologs in mouse, rat, and dog genomes. This requirement reduces the number of duplicate-free human genes for the analysis by ~25% (from 4,857 to 3,681). To maintain the statistical power and keep our dataset representative of the whole genome, we generated a tandem-duplicate-free dataset, which contains 7,577 human genes that have expression data and have orthologs in the other three mammalian genomes. This dataset is larger than the above duplicate-free dataset because we now allow duplicate genes that are located on different chromosomes.

4.3.4 Expression-profile similarity between linked genes

Following Gu et al. (2002), we measured the level of co-expression between two linked genes (say A and B) by $ln[(1+R)/(1-R)]$, where R is Pearson's correlation coefficient

of signal intensity *S* across all the tissues examined. Higher *ln*[(1+*R*)/(1-*R*)] indicates a higher level of co-expression. Using *R* instead of *ln*[(1+*R*)/(1-*R*)] does not change any of our results qualitatively. The chromosomal distance (*D*) between linked genes was defined by the distance (in nucleotides) between the transcription starting sites of the two genes, as annotated by Ensembl.

In Figs. 3 and S3, the size of each bin was fixed to a certain value. In Fig. 2, because the genomic distance *D* was log-transformed when the linear regression was applied, we gradually increase the bin size as *D* increases to avoid the overrepresentation of data points with large *D*. The size of the *n*th bin is $10^6 \times 2^{(99+n)/100}$ nucleotides. That is, the *n*th bin represents the group of linked genes with *D* values ranging from $10^6 \times \sum_{i=1}^{n-1} 2^{(99+i)/100}$ to $10^6 \times \sum_{i=1}^{n} 2^{(99+i)/100}$, except for the first bin which is with *D* from 1 to 2×10^6. Use of other bin sizes did not change our results qualitatively.

4.3.5 Evolutionary conservation of linkage

When a gene pair is linked in both human and dog genomes, we regard the linkage to be old (or ancestral). Here and elsewhere in this paper, linkage means that two genes are located in the same chromosome. Although it is possible that two previously unlinked genes became linked in human and dog independently, such events have low probabilities and can be ignored. We analyzed the subset of human gene pairs with old linkages. Within this subset, if the linkage for a gene pair is maintained in both mouse and rat genomes, the linkage is said to be "conserved"; otherwise, it is non-conserved, meaning that the linkage is lost in one or both rodents.

4.4 RESULTS

4.4.1 Co-expression of distantly linked human genes

It is important to first examine whether the phenomenon of co-expression of linked genes exists for both closely and distantly linked genes, as such knowledge can help understand the relative importance of different molecular mechanisms responsible for the phenomenon (Hurst, Pal, and Lercher 2004). Some studies have attempted to address this question by examining the on/off expressional status of linked genes (Lercher, Urrutia, and Hurst 2002; Semon and Duret 2006), while other studies examined the correlation of across-tissue expression-profiles of adjacent genes (Hurst, Williams, and Pal 2002; Singer et al. 2005). Adjacent genes are linked genes without any other genes in between. Since chromosomal rearrangements between sex chromosomes and autosomes are rare and sex-linked genes have special functions and expression profiles (Lahn, Pearson, and Jegalian 2001; Wang et al. 2001), here we limit our analyses to autosomal genes. From the 4,857 duplicate-free human autosomal genes (see Materials and Methods), we obtained 4,835 adjacent gene pairs. Let D be the distance in nucleotides between a pair of linked genes in a chromosome. We find a significant correlation between log$D$ and the level of co-expression ($ln[(1+R)/(1-R)]$, see Materials and Methods) (Pearson's correlation coefficient $r = -0.1385$, $P < 10^{-21}$; Spearman's correlation coefficient $\rho = -0.1364$, $P < 10^{-20}$), indicating that closer adjacent human genes have higher similarity in spatial expression-profiles. Because microarray data are known to be noisy, to reduce the effect of stochastic background noise, we group linked genes with similar D and calculate average $ln[(1+R)/(1-R)]$ for each group. The aforementioned pattern can be seen more clearly with binned data (Figure 4.1A). In comparison, the human genome with permutated expression-profiles, which is generated by

randomly assigning gene names to the real expression-profiles (of the 4835 duplicate-free genes), shows no obvious pattern between log$D$ and ln[(1+R)/(1-R)] (Figure 4.1B). Our observations are consistent with previous studies in the yeast (Hurst, Williams, and Pal 2002).

The power to decipher the effect of linkage on gene co-expression is limited if only adjacent genes are analyzed, because there are few adjacent genes with large intervening distances (e.g. 713 adjacent gene pairs with D > 1 Mb and 10 with D > 10 Mb in our dataset). We thus analyze pairs of linked genes, without requiring them to be adjacent to each other. From 4,857 duplicate-free human autosomal genes (see above), we obtain 518,133 linked gene pairs with the genomic distances ranging up to one hundred megabases. We then group the gene pairs according to their D values (see Materials and Methods) and calculate the average ln[(1+R)/(1-R)] for each group. We observe a strong negative correlation between log$D$ and ln[(1+R)/(1-R)] (Pearson's r = -0.7121, $P < 10^{-80}$; Spearman's ρ = -0.6227, $P < 10^{-56}$; Fig. 2). On the contrary, the genome with permutated expression-profiles shows no correlation (Pearson's r = 0.0198, P = 0.654; Spearman's ρ = -0.0700, P = 0.1122; Figure A.8). Since the correlation observed in the real human genome is computed from the data points with D varying from 10 kilobases to 100 megabases (Figure 4.2), it is possible that the correlation is solely caused by the data points with small D values (e.g., < 1 Mb). To examine this possibility, we divided our data into five categories based on the value of D: <1 Mb, 1-5 Mb, 5-25 Mb, 25-50 Mb and 50-100 Mb. The negative correlation between log$D$ and ln[(1+R)/(1-R)] is significant in nearly every category for the real genome (Table 4.1). To know the chance probability of observing these correlations, we generate 1,000 permutated genomes by randomly swapping gene names of the expression profiles. The chance probability is the frequency of the observed correlations in randomly permutated

genomes that are more negative than the correlation observed in the real genome. The result shows that the probabilities are <0.001 in categories <1Mb, 1-5Mb and 5-25Mb (Table 4.1), indicating that the phenomenon of co-expression of linked genes extends to a distance of tens of megabases in humans, which can harbor several hundred genes. In addition to *D*, we also measure the distance between two linked genes by the number (*N*) of intervening genes between them. Consistent with Figure 4.2, the correlation between *N* in $\log_2$ scale and $ln[(1+R)/(1-R)]$ is significantly negative (Figure A.9), indicating that our observation does not depend on how the distance is measured and that linked genes with >100 intervening genes are still significantly co-expressed.

To examine whether the phenomenon of long-range gene co-expression is universal in mammals, we apply the same method for generating Table 1 to the mouse data (see Materials and Methods). Although the correlation between log*D* and $ln[(1+R)/(1-R)]$ is significant when $D < 1$ Mb and 5-25Mb, the negative correlations do not exist for the groups of 1-5Mb, 25-50Mb and 50-100Mb in mouse (Table A.5).

4.4.2 Weaker evolutionary conservation of linkage between genes of higher co-expression

If the long-range co-expression of linked genes in humans is an outcome of adaptive evolution, the gene order in a large part of the human genome must have been highly organized and evolutionarily preserved. An important test of the hypothesis of functional relevance and adaptive value of co-expression of linked genes is to measure the evolutionary conservation of linkage. If co-expression of linked genes is favored by natural selection, the linkage should be maintained during evolution. If co-expression of linked genes is a neutral

phenomenon without functional consequences, no difference in conservation of linkage is expected between gene pairs with high levels of co-expression and those with low levels of co-expression. If co-expression of linked genes is detrimental, the linkage of highly co-expressed genes should be broken more often during evolution than that of weakly co-expressed genes. To test these hypotheses, we utilize the tandem-duplicate-free 7,577 human genes that have orthologs in each of the mouse, rat, and dog genomes (see Materials and Methods). Based on the mammalian phylogeny shown in Fig. 3A (Springer et al. 2003; Murphy, Pevzner, and O'Brien 2004; Kriegs et al. 2006; Nishihara, Hasegawa, and Okada 2006), we infer that two linked human genes were also linked in the common ancestor of primates, rodents, and carnivores, if their orthologs are linked in the dog genome (see Materials and Methods). Note that although some authors believe that primates and carnivores are more closely related to each other than each is to rodents (Cannarozzi, Schneider, and Gonnet 2006), the phylogeny we use here has been well established by analyses of both irreversible genomic events (Kriegs et al. 2006; Nishihara, Hasegawa, and Okada 2006) and DNA sequences from many taxa (Springer et al. 2003; Murphy, Pevzner, and O'Brien 2004), and thus is much more reliable than results based on DNA sequences from only a few taxa. In the present study, we only investigate ancestrally linked gene pairs because only these genes can be used to unambiguously determine the breakage of linkage (Table 4.3A). Because rodent genomes have gone through extensive rearrangements during evolution (Bourque, Pevzner, and Tesler 2004; Mullins and Mullins 2004), current organizations of mouse and rat genomes help divide these ancestrally linked genes into two groups: genes with conserved linkage and genes with non-conserved linkage. We then compare the level of co-expression between gene pairs with conserved linkage and those

with non-conserved linkage. Since genomic distance *D* influences expression similarity (Table 4.1 and 4.2), we control the effect of *D* by grouping genes with similar *D* values, and then compare average $ln[(1+R)/(1-R)]$ values of the conservatively linked genes and non-conservatively linked genes within each group. The results show that, for nearly every *D* range, non-conservatively linked human genes have a higher degree of co-expression than conservatively linked human genes (Table 4.3B and 4.3C). This finding is inconsistent with the adaptive model (Hurst, Williams, and Pal 2002; Singer et al. 2005) and the neutral model (Semon and Duret 2006), but is predicted by our hypothesis that co-expression of linked genes is generally detrimental and disfavored by natural selection. We also compare the expression similarity between conservatively and non-conservatively linked gene pairs when we define non-conservation by a loss of linkage in primates, instead of rodents. The results (Figure A.10) are similar to those in Figure 4.3B and 4.3C, suggesting that the phenomenon of weaker evolutionary conservation of linkage between genes of higher co-expression is not unique to one particular mammalian lineage, but is likely to be generally true in mammals.

One interesting question is whether the selection against co-expression (or interference) only acts on weakly to moderately co-expressed linked genes but not on strongly co-expressed linked genes. To define strongly co-expressed genes, we plotted the distribution of $ln[(1+R)/(1-R)]$ for all 1,521,714 linked gene pairs (from 7,577 tandem duplicate-free genes used in Figure 4.3A-C), and considered linked genes with $ln[(1+R)/(1-R)]$ values falling within the top 5% of the distribution (Figure 4.3D) to be strongly co-expressed. Interestingly, we found that the proportion of strongly co-expressed gene pairs is lower among those with conserved linkage than with non-conserved lineage (Figure 4.3E and

4.3F), suggesting natural selection against the conservation of linkage of strongly co-expressed gene pairs.

Our transcriptional interference hypothesis predicts that the breakage of linkage between two genes would reduce the degree of their co-expression. We examine the difference between the expression-profile similarity of human linked gene pairs and that of their mouse orthologs, by using 26 human-mouse common tissues. The full list of these 26 tissues can be found in a previous study (Liao and Zhang 2006a). Since co-expression of linked genes is much weaker in mouse than in human (Table A.5), there is a general trend of reduction in expression-profile similarity between a gene pair in mouse compared to that in human (Figure 4.4). However, the reduction is greater for the gene pairs that experienced inter-chromosomal rearrangements than those that did not (Figure 4.4). This finding is consistent with the hypothesis that chromosomal rearrangement helps reduce transcriptional interference.

Some authors suggested that reduced recombination can ensure the physical proximity of linked genes (Pal and Hurst 2003; Poyatos and Hurst 2006). Therefore, one expects to observe lower recombination rates between highly co-expressed genes than between poorly co-expressed genes, if co-expression of linked genes is beneficial. However, our analysis of the human genome shows that highly co-expressed linked genes actually have higher recombination rates (cM/Mb) than poorly co-expressed linked genes (Figure A.11). Although recombination rate and chromosomal rearrangement may not be independent from each other (Akhunov et al. 2003; Lindsay et al. 2006), our observation again argues against the adaptive model and neutral model, but is consistent with our hypothesis that co-expression of linked genes is detrimental.

4.5 DISCUSSION

There are generally three molecular mechanisms that could cause the co-expression of linked genes (Hurst, Pal, and Lercher 2004). At the primary level, *cis*-acting elements directly affect the transcription of neighboring genes (Cho et al. 1998; Kruglyak and Tang 2000). This mechanism will only affect genes within a few kilobases of one another. At the secondary level, histone modifications spread from a locus control region to co-suppress the transcriptional activities of several linked genes until reaching boundary elements (Labrador and Corces 2002). This type of co-regulation affects regions of up to a few hundred kilobases. At the tertiary level, transcriptional co-regulation can happen in two ways. First, genes with certain *cis*-acting elements can come together to form the node of chromatin loops during transcription; such special formation of aggregated *cis*-elements is called the active chromatin hub (ACH); genes close to the ACH are accessible to transcription, whereas genes looping out are inaccessible (de Laat and Grosveld 2003). Second, arrangement of chromatin in compact chromosome territories can affect transcription; transcription is largely restricted to territory surfaces but suppressed within the interior (Cremer and Cremer 2001). In both of these tertiary-level regulations, effects are expected to range up to several megabases.

In the present work, we first report the phenomenon of very-long-range (up to tens of megabases) co-expression of linked genes in the human genome. Although this result might suggest the importance of tertiary-level transcriptional regulations in humans, to our knowledge, there is no mechanism that has been demonstrated to regulate co-expression of linked genes at such large distances. Is it possible that our observation is merely an artifact?

One potential caveat is the design of the microarray chip that is used to generate the gene expression data. For example, yeast cDNA arrays are designed with the probes printed in genomic order and it has been suggested that previously observed periodicity of expression patterns of genes located in a chromosome (Cho et al. 1998; Cohen et al. 2000; Kruglyak and Tang 2000) is due to the spatial order of probes on the array (Lercher and Hurst 2006). Since the expression data used here is produced from oligonucleotide microarrays for which the probe positions appear random (Su et al. 2004), the spatial bias occurred in the yeast cDNA array cannot explain our observation. Another possible caveat is the potential unequal levels of co-expression of linked genes on different chromosomes. If the level of co-expression is higher in small chromosomes than in large chromosomes for a given *D*, the results of Figure 4.2 and Table 4.1 may be generated simply by the bias of sampling more gene pairs with large *D* from large chromosomes. However, we do not find any correlation between the level of co-expression and chromosomal size when controlled for *D* (Figure A.12 and A.13). Moreover, the negative correlation between the level of gene co-expression and physical distance within a single chromosome is similar to the genome-wide pattern (Figure A.14). It is worth mentioning that one yeast study proposed that the seemingly long-range co-expression of linked genes is perhaps due to similar expression patterns of genes in subtelomeric regions (Lercher and Hurst 2006). We examine this hypothesis by reproducing Figure 4.2 after removing human genes in subtelomeric regions (<5 Mb from chromosomal ends). The result shows a virtually identical correlation between log*D* and $ln[(1+R)/(1-R)]$ (Pearson's $r = -0.7123$, $P < 10^{-80}$; Spearman's $\rho = -0.6408$, $P < 10^{-60}$) as in Figure 4.2, suggesting that our results are not due to special genes in subtelomeric regions. We conclude that the long-range co-expression of human linked genes is real, although the underlying

molecular mechanism remained to be investigated. It should be noted that our result does not imply that the primary and secondary levels of gene regulation are unimportant. Rather, the patterns observed in Figure 4.1 and 4.2 suggest the existence of these two levels of regulation as well.

Contrary to the hypothesis that co-expressed gene clusters correspond to large chromatin domains (Hurst, Williams, and Pal 2002; Roy et al. 2002; Hurst, Pal, and Lercher 2004; Sproul, Gilbert, and Bickmore 2005), a recent study showed that co-expression of mammalian genes is mainly due to the co-regulation of two genes by shared promoters (Semon and Duret 2006). Our result favors the hypothesis of gene co-regulation by large domains, which is consistent with the discovery in yeast (Lercher and Hurst 2006). Different from our approach, Semon and Duret (Semon and Duret 2006) followed the method used in Lercher et al. (Lercher, Urrutia, and Hurst 2002) to measure the expression-profile similarity of two linked genes by calculating how often they are simultaneously "turned on". One explanation for the inconsistency of our results with that of Semon and Duret (Semon and Duret 2006) is the fact that transcriptional background only affects the relative gene expression levels across different tissues, but not a change of the on/off status of a gene in a particular condition. In such cases, it is more sensitive to measure co-expression of two genes by Pearson's correlation coefficient R. Other drawbacks of using the on/off status to measure expression-profile similarities from microarray data have been thoroughly discussed in an earlier paper (Liao and Zhang 2006a).

Previous investigators have used evolutionary conservation of linkage to study the potential adaptive value of linkage of co-expressed genes, but they did not use outgroups to separate the formation of new linkages from the breakage of old linkages (Hurst, Williams,

and Pal 2002; Singer et al. 2005; Semon and Duret 2006). Hence, if a pair of highly co-expressed genes is observed to be linked in one genome (species A) but not in another (species B), it is often interpreted as a breakage of linkage in species B. In fact, this observation could also be due to the formation of the linkage in species A since the separation of the two species. These two scenarios cannot be differentiated without the use of an outgroup genome. In the present study, we use the dog as an outgroup to identify those gene pairs that were ancestrally linked in the common ancestor of primates, rodents, and carnivores. We found more inter-chromosomal rearrangements during rodent evolution for gene pairs with high co-expression in humans than those with low co-expression (Figure 4.3). Therefore, co-expression of linked genes appears to be disfavored by natural selection. To examine whether using an outgroup would drastically change the conclusion of previous studies that supported the adaptive model, we repeated the analyses of Singer et al. (Singer et al. 2005) by counting the inter-chromosomal breakages within clusters (see Figure 4C in (Singer et al. 2005)) that occurred in the mouse lineage after the divergence of primates and rodents. The new result (Figure A.15) is opposite of Singer et al.'s result and becomes consistent with our findings in Figure 4.3.

Our observations suggest no adaptive value for clustering of co-expressed genes in the human genome in general. Rather, linked genes are co-expressed simply because they share a similar transcriptional background. The existence of large genomic regions with a similar transcriptional background implies that many mammalian genes may never reach their optimal expression-profiles because of the interference of the surrounding genomic environment. It should be noted that some authors proposed that the linkage of co-expressed genes may represent lineage-specific transient adaptations (Ranz et al. 2007; Poyatos and

Hurst 2007). While this scenario remains possible, it is extremely hard to test by comparative approaches. Furthermore, this scenario is not contradictory to our finding that co-expression of linked genes is generally deleterious over long-term evolution.

Note that we do not suggest that eukaryotic gene order is completely random. Apart from the gene clusters formed by gene duplication or operons (Lercher, Blumenthal, and Hurst 2003; Hurst, Pal, and Lercher 2004), many clusters of functionally related genes do exist, such as clusters of genes encoding organelle-related proteins (Lefai et al. 2000; Elo et al. 2003; Alexeyenko et al. 2006) and genes encoding proteins in the same protein complex (Teichmann and Veitia 2004). However, it should be noted that some of these clusters actually do not show high degree of gene co-expression (Alexeyenko et al. 2006). Together with our finding, it is clear that the phenomenon of co-expression and similar function of linked genes should be considered separately. A recent study showed that gene expression-profile corresponds poorly to gene function (Yanai et al. 2006). Apparently, there are factors other than gene function that determine a gene's expression. Because evolutionary changes of gene expression may play a more significant role than changes of protein sequence in phenotypic evolution (King and Wilson 1975; Carroll 2005), identifying such factors is of fundamental importance to our understanding of evolution. Our result implies that a change in gene location can facilitate expression evolution, which is similar to what was previously known as the positional effect (Festenstein et al. 1996; Milot et al. 1996; Kleinjan and van Heyningen 1998).

Our hypothesis that co-expression of linked genes is detrimental raises an important question. That is, if such co-expression is deleterious, how can it be fixed in the first place? Here, we propose a model to explain this seemingly dilemmatic phenomenon. We propose

that although co-expression of linked genes is generally detrimental, the "mutation" that generates co-expression as a byproduct may initially be advantageous. Figure 4.5 shows an example explaining this model. For simplicity, only two genes, A and B, are shown. Initially, A and B are linked but with distinct expression patterns (Figure 4.5A). However, the expression of B is not optimized. When a mutation occurs to establish a transcriptional background for the two genes, they become co-expressed. This mutation makes the expression pattern of B closer to its optimal, while the co-expression makes the expression pattern of A deviate from its optimal (Figure 4.5B). The overall fitness gain may still be positive for these changes and the mutation could be fixed by either positive selection or drift. However, because the expression of A is suboptimal, subsequent breakage of the A-B linkage and move of A to another genomic location may be advantageous (Figure 4.5C). It is possible that many genes are involved in a similar evolutionary process as shown in Figure 4.5, since the mechanism creating the transcriptional background have long-range effects. The above verbal model lacks many quantitative details, because the molecular mechanism responsible for co-regulation of linked genes is poorly known. In the future, when the molecular mechanism of co-regulation is better understood, it would be interesting to study the feasibility of the above model using population genetic analysis and computer simulation.

Chromosomal rearrangement is just one way to remove the transcriptional interference of linked genes (Figure 4.3). Other mechanisms, such as the increase of intergenic distance (Byrnes, Morris, and Li 2006) and establishment of insulators (Bell, West, and Felsenfeld 2001), have also been reported. We found the phenomenon of long-range co-expression of linked genes to be much more prominent in human than in mouse (Table 4.1 and Table A.5), consistent with the earlier observation that short-range co-expression is also

more prominent in human than in mouse (Singer et al. 2005). A simple explanation of the human-mouse difference is that the mouse gene expression data had higher background noise compared to the human data, resulting in weaker co-expression signals that are identifiable by our method. However, it is beyond our ability to confirm this explanation. It is possible that the high rate of chromosomal rearrangement in rodents is in part responsible for the less significant co-expression of linked genes in the mouse genome, because rearranged mouse orthologs of human linked genes have a greater reduction in expression-profile similarity than non-rearranged mouse orthologs (Figure 4.4). However, because large conserved syntenic blocks (>50 megabases) still exist between human and mouse and the total number of syntenic blocks is no more then 400 (Waterston et al. 2002; Bourque, Pevzner, and Tesler 2004; Liao et al. 2004), chromosomal rearrangements in rodents are unlikely to be sufficient to completely "scramble" the mouse genome. Hence, assuming no quality difference in either genomic sequence or gene expression data between human and mouse, we cannot exclude the possibility that other mechanisms exist in rodents to alleviate transcriptional interference of linked genes. As the population size is larger for rodent species than for primate species, natural selection promoting the reduction of transcriptional interference may be more efficient in rodents than in primates. It would be interesting to test this hypothesis in the future.

In conclusion, our observations presented in the present study are consistent with neither the adaptive nor the neutral model. The results support our hypothesis that co-expression of linked genes in the human genome is a form of deleterious transcriptional interference. Because all genes are located in the neighborhood of other genes, such interference may be mechanistically inevitable. As a consequence, the expression-profile of

a gene may never be optimized in evolution. Rather, transcriptional interference may be the source creating instability and dynamics of the mammalian gene order. In light of this finding, it will be of great interest to identify those few genes that are tightly linked across a large number of mammals or vertebrates, as such exceptional incidences of conserved linkage (e.g., Hox clusters) likely indicate gene co-regulations that are beneficial to the organisms.

4.6 ACKNOWLEDGMENTS

We thank Xionglei He, Wendy Grus, Ondrej Podlaha, Zhi Wang, and Patricia Wittkopp for valuable comments. This work was supported by research grants from University of Michigan Center for Computational Medicine and Biology and National Institutes of Health to J.Z.

Figure 4.1 Low co-expression for closer human adjacent genes. Expression-profile similarities between adjacent human genes, measured by *ln*[(1+*R*)/(1-*R*)], is negatively correlated with log*D*, their $\log_{10}$-transformed genomic distance in nucleotides, in **(A)** the real human genome, but not in **(B)** the permutated human genome. Average *ln*[(1+R)/(1-R)] (± standard error) are shown for each group of adjacent genes categorized by log*D*. The number of gene pairs per category is 213, 488, 966, 1343, 1111, and 714, respectively, for the six categories.

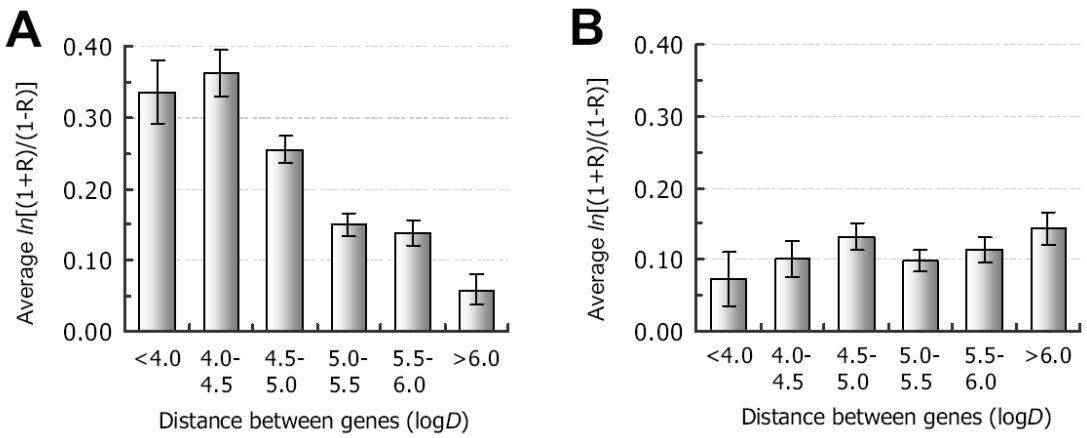


Figure 4.2 Low co-expression for closer human linked genes. Linear regression of average expression-profile similarity of linked genes, measured by *ln*[(1+*R*)/(1-*R*)], versus their $\log_{10}$-transformed genomic distance in nucleotides (log*D*), where *D* is set to be the median of each X-axis bin. In the real human genome, average *ln*[(1+*R*)/(1-*R*)] is strongly negatively correlated with log*D***.** The bin size ranges from 20 kilobases (the 1st bin) to ~ 715 kilobases (the last bin) (see Materials and Methods for details on bin sizes). The figure is further divided into five areas by gray shading. These fives areas are <1Mb, 1-5Mb, 5-25M, 25-50Mb and 50-100Mb, respectively. The correlations between *ln*[(1+*R*)/(1-*R*)] and log*D* of these five areas are shown in Table 4.1.

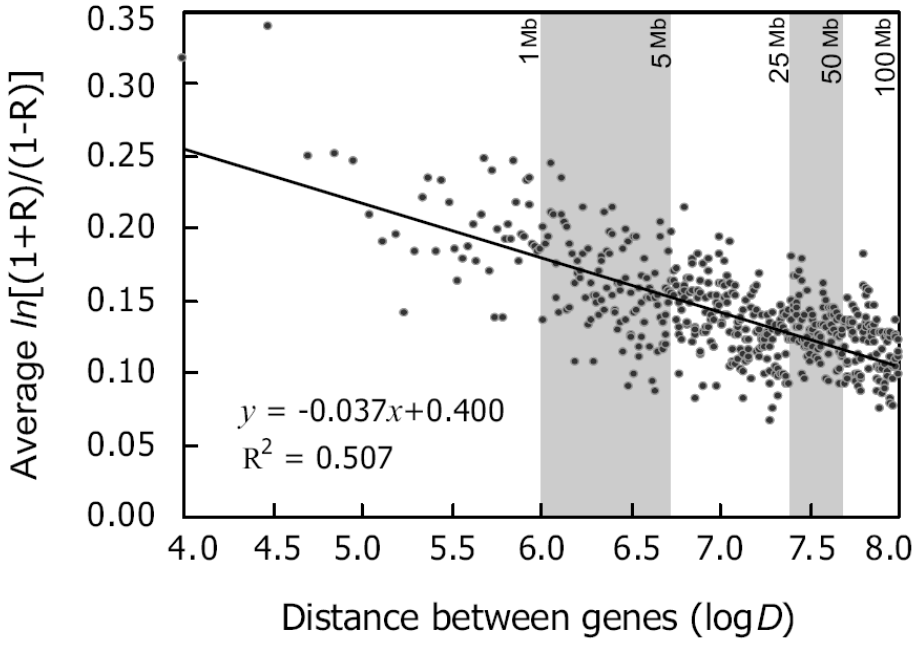


Figure 4.3 Linked human genes with non-conserved linkage have higher expression-profile similarity than those with conserved linkage. (A) Phylogeny of human, rat, mouse, and dog. Only human ancestrally linked genes, determined by the linkage in dog, are included in the analysis. Pink and green bars represent two genes. Average $ln[(1+R)/(1-R)]$ (± standard error) for genes with conserved linkage and genes with non-conserved linkage, at **(B)** short physical distances and at (**C**) short physical distances. **(D)** Distribution of $ln[(1+R)/(1-R)]$ for all linked (duplicate-free) gene pairs. Strongly co-expressed linked genes are those that fall in the 5% right-tail of the distribution. They have a minimal $ln[(1+R)/(1-R)]$ of 1.25. Average $ln[(1+R)/(1-R)]$ (± standard error) for strongly co-expressed genes with conserved linkage and genes with non-conserved linkage, at **(E)** short physical distances and at **(F)** short physical distances. *P* values (paired *t*-test) for the hypothesis of no difference in mean expression-profile similarity between genes with conserved linkage and those with non-conserved linkage are 6.37×10^{-2}, 7.91×10^{-6}, 9.70×10^{-3}, and 2.99×10^{-4} for **(B), (C), (E)** and **(F),** respectively.

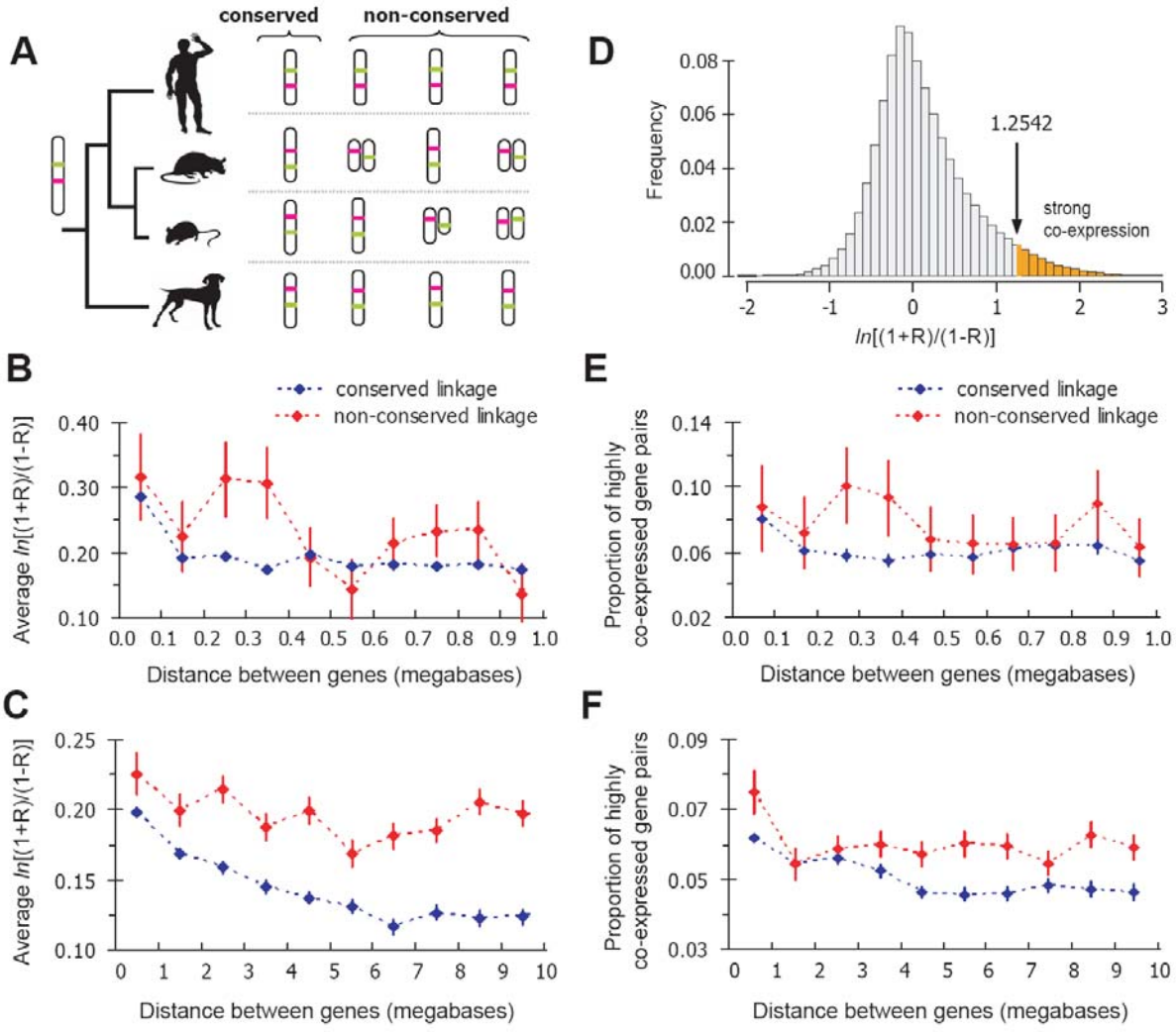


Figure 4.4 Co-expression of linked genes is reduced by inter-chromosomal rearrangements. Only human linked genes that are ancestrally linked, determined by the linkage in dog, are included in the analysis. Mouse orthologs of these ancestrally linked genes can be either linked on the same chromosome (white bars) or separated on different chromosomes (black bars). Y-axis shows the difference in expression-profile similarity (± standard error), measured by $ln[(1+R)/(1-R)]$, of two human linked genes and that of their mouse orthologs. The *P* value (paired *t*-test) for the hypothesis of no difference in average reduction of expression-profile similarity between the two groups of genes is 7.83×10^{-4}.

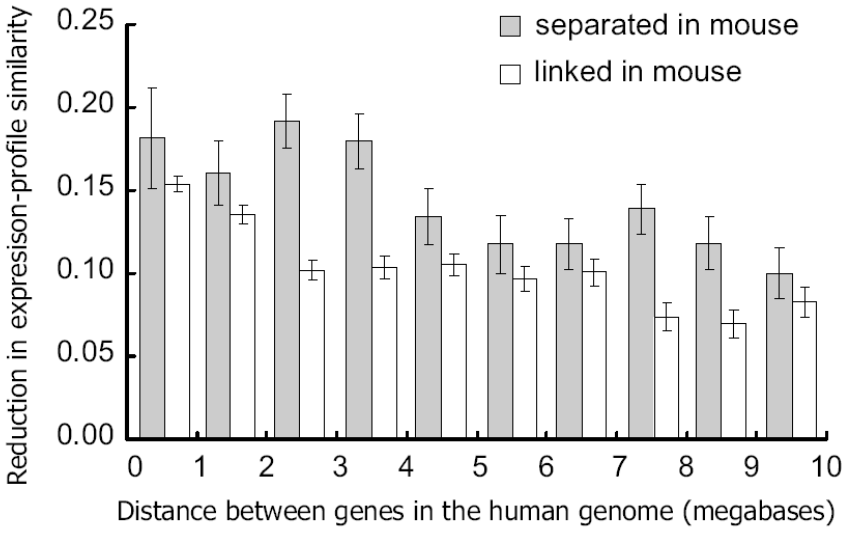


Figure 4.5 The birth and breakdown of co-expression of linked genes. (A) The initial expression status of gene A and B. Gene A and B are linked but not co-expressed. The expression-profiles of A and B are shown above the boxes representing the genes. Solid lines represent the current expression-profiles, whereas dashed gray lines represent the optimal expression-profiles for a gene to carry its functions. I, II, II, and IV represent different conditions or tissues. **(B)** The birth of the co-expression of gene A and gene B. The establishment of the transcriptional background suppresses the gene expression under condition III. Pink arrows show the directions of suppression from the initiation site. This mutation drives the expression-profile of B closer to, but makes that of A away from, its optimal expression-profile. This mutation also causes co-expression of A and B as a byproduct. Although this mutation is detrimental to the function of A, the net fitness gain for the organism is positive, and thus the mutation establishing the transcriptional background can be fixed. The contribution to an organism's fitness gain by the expression-profile change is marked below the box representing the gene. **(C)** The breakdown of the co-expression of A and B. A chromosomal rearrangement disrupts the linkage between A and B, terminating the interference of transcriptional background on the expression of A. A and B are no longer co-expressed. Because the rearrangement increases the overall fitness, this mutation can be fixed.

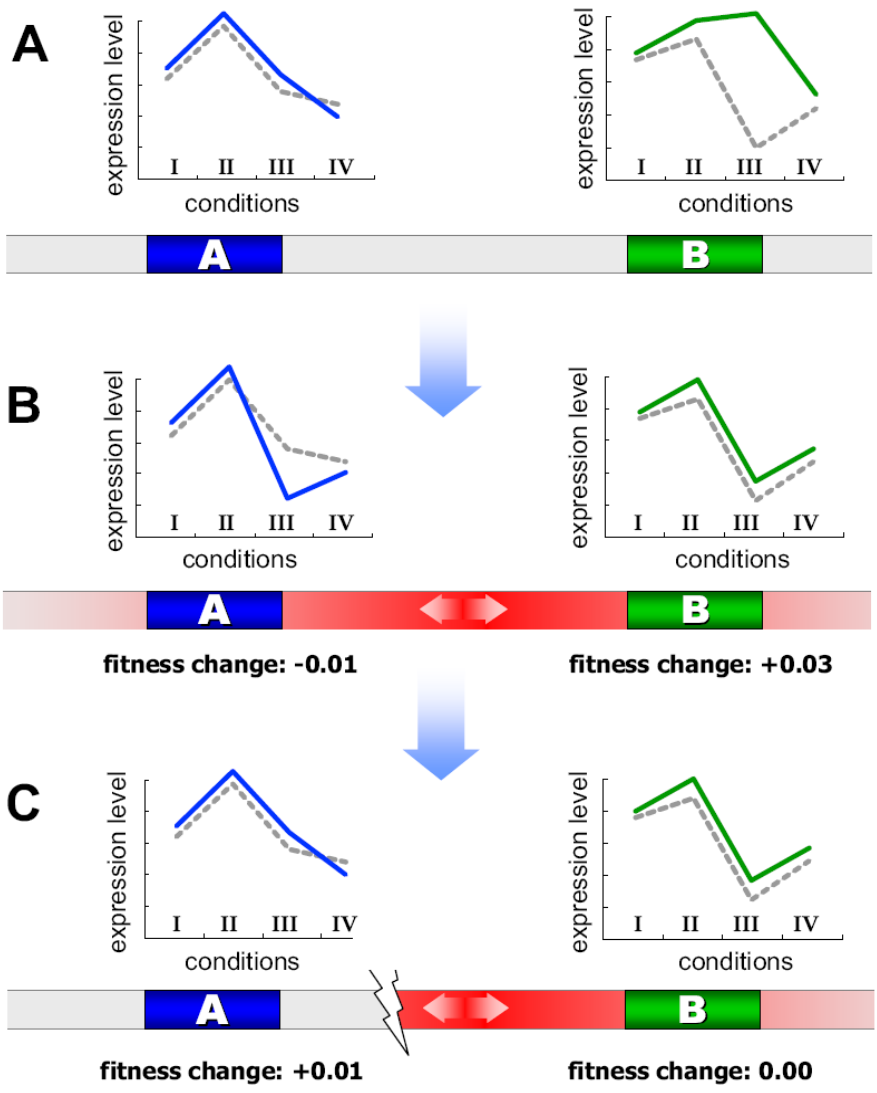


Table 4.1 Correlation between chromosomal distance (log*D*) and average expression-profile similarity, measured by *ln*[(1+*R*)/(1-*R*)], between human linked gene pairs. Correlations are calculated from subsets of gene pairs with different ranges of genomic distances (*D*). The chance probability of observing a correlation as strong as observed is determined from 1,000 permutated genomes. The original data points for the real human genome are shown in Fig. 2.

Genomic distance (*D*)	Pearson's *r*	Chance probability
<1Mb	-0.5808	< 0.001
1-5Mb	-0.4523	< 0.001
5-25Mb	-0.4416	< 0.001
25-50Mb	-0.2693	0.069
50-100Mb	-0.2391	0.119

4.7 LITERATURE CITED

Akhunov, ED, AR Akhunova, AM Linkiewicz et al. 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc Natl Acad Sci U S A **100**:10836-10841.

Alexeyenko, A, AH Millar, J Whelan, and EL Sonnhammer. 2006. Chromosomal clustering of nuclear genes encoding mitochondrial and chloroplast proteins in Arabidopsis. Trends Genet **22**:589-593.

Bailey, JA, R Baertsch, WJ Kent, D Haussler, and EE Eichler. 2004. Hotspots of mammalian chromosomal evolution. Genome Biol **5**:R23.

Bell, AC, AG West, and G Felsenfeld. 2001. Insulators and boundaries: versatile regulatory elements in the eukaryotic. Science **291**:447-450.

Bourque, G, PA Pevzner, and G Tesler. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res **14**:507-516.

Boutanaev, AM, AI Kalmykova, YY Shevelyov, and DI Nurminsky. 2002. Large clusters of co-expressed genes in the Drosophila genome. Nature **420**:666-669.

Byrnes, JK, GP Morris, and WH Li. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol Biol Evol **23**:1136-1143.

Cajiao, I, A Zhang, EJ Yoo, NE Cooke, and SA Liebhaber. 2004. Bystander gene activation by a locus control region. Embo J **23**:3854-3863.

Cannarozzi, GM, A Schneider, and G Gonnet. 2006. A Phylogenomic Study of Human, Dog and Mouse. PLoS Comput Biol:, in press.

Caron, H, B van Schaik, M van der Mee et al. 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291**:1289-1292.

Carroll, SB. 2005. Evolution at two levels: on genes and form. PLoS Biol **3**:e245.

Cho, RJ, MJ Campbell, EA Winzeler et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell **2**:65-73.

Cohen, BA, RD Mitra, JD Hughes, and GM Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet **26**:183-186.

Cremer, T, and C Cremer. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet **2**:292-301.

de Laat, W, and F Grosveld. 2003. Spatial organization of gene expression: the active chromatin hub. Chromosome Res **11**:447-459.

Denver, DR, K Morris, JT Streelman, SK Kim, M Lynch, and WK Thomas. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet **37**:544-548.

Elo, A, A Lyznik, DO Gonzalez, SD Kachman, and SA Mackenzie. 2003. Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the Arabidopsis genome. Plant Cell **15**:1619-1631.

Eszterhas, SK, EE Bouhassira, DI Martin, and S Fiering. 2002. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol Cell Biol **22**:469-479.

Festenstein, R, M Tolaini, P Corbella, C Mamalaki, J Parrington, M Fox, A Miliou, M Jones, and D Kioussis. 1996. Locus control region function and heterochromatin-induced position effect variegation. Science **271**:1123-1125.

Fischer, G, EP Rocha, F Brunet, M Vergassola, and B Dujon. 2006. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genet **2**:e32.

Fukuoka, Y, H Inaoka, and IS Kohane. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics **5**:4.

Hubbell, E, WM Liu, and R Mei. 2002. Robust estimators for expression analysis. Bioinformatics **18**:1585-1592.

Hurst, LD, C Pal, and MJ Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet **5**:299-310.

Hurst, LD, EJ Williams, and C Pal. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet **18**:604-606.

Huynen, MA, B Snel, and P Bork. 2001. Inversions and the dynamics of eukaryotic gene order. Trends Genet **17**:304-306.

Jordan, IK, L Marino-Ramirez, and EV Koonin. 2005. Evolutionary significance of gene expression divergence. Gene **345**:119-126.

Kalmykova, AI, DI Nurminsky, DV Ryzhov, and YY Shevelyov. 2005. Regulated chromatin domain comprising cluster of co-expressed genes in Drosophila melanogaster. Nucleic Acids Res **33**:1435-1444.

Khaitovich, P, I Hellmann, W Enard, K Nowick, M Leinweber, H Franz, G Weiss, M Lachmann, and S Paabo. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science **309**:1850-1854.

Khaitovich, P, G Weiss, M Lachmann, I Hellmann, W Enard, B Muetzel, U Wirkner, W Ansorge, and S Paabo. 2004. A neutral model of transcriptome evolution. PLoS Biol **2**:682-689.

King, MC, and AC Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.

Kleinjan, DJ, and V van Heyningen. 1998. Position effect in human genetic disease. Hum Mol Genet **7**:1611-1618.

Kriegs, JO, G Churakov, M Kiefmann, U Jordan, J Brosius, and J Schmitz. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol **4**:e91.

Kruglyak, S, and H Tang. 2000. Regulation of adjacent yeast genes. Trends Genet **16**:109-111.

Labrador, M, and VG Corces. 2002. Setting the boundaries of chromatin domains and nuclear organization. Cell **111**:151-154.

Lahn, BT, NM Pearson, and K Jegalian. 2001. The human Y chromosome, in the light of evolution. Nat Rev Genet **2**:207-216.

Lawrence, J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev **9**:642-648.

Lee, JM, and EL Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res **13**:875-882.

Lefai, E, MA Fernandez-Moreno, LS Kaguni, and R Garesse. 2000. The highly compact structure of the mitochondrial DNA polymerase genomic region of Drosophila melanogaster: functional and evolutionary implications. Insect Mol Biol **9**:315-322.

Lercher, MJ, T Blumenthal, and LD Hurst. 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res **13**:238-243.

Lercher, MJ, and LD Hurst. 2006. Co-expressed yeast genes cluster over a long range but are not regularly spaced. J Mol Biol **359**:825-831.

Lercher, MJ, AO Urrutia, and LD Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet **31**:180-183.

Liao, BY, YJ Chang, JM Ho, and MJ Hwang. 2004. The UniMarker (UM) method for synteny mapping of large genomes. Bioinformatics **20**:3156-3165.

Liao, BY, and J Zhang. 2006a. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol **23**:1119-1128.

Liao, BY, and J Zhang. 2006b. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol **23**:530-540.

Lindsay, SJ, M Khajavi, JR Lupski, and ME Hurles. 2006. A chromosomal rearrangement hotspot can be identified from population genetic variation and is coincident with a hotspot for allelic recombination. Am J Hum Genet **79**:890-902.

Megy, K, S Audic, and JM Claverie. 2003. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. Genome Biol **4**:P1.

Miller, MA, AD Cutter, I Yamamoto, S Ward, and D Greenstein. 2004. Clustered organization of reproductive genes in the C. elegans genome. Curr Biol **14**:1284-1290.

Milot, E, J Strouboulis, T Trimborn et al. 1996. Heterochromatin effects on the frequency and duration of LCR-mediated gene transcription. Cell **87**:105-114.

Mullins, LJ, and JJ Mullins. 2004. Insights from the rat genome sequence. Genome Biol **5**:221.

Murphy, WJ, PA Pevzner, and SJ O'Brien. 2004. Mammalian phylogenomics comes of age. Trends Genet **20**:631-639.

Nishihara, H, M Hasegawa, and N Okada. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc Natl Acad Sci U S A **103**:9929-9934.

Nuzhdin, SV, ML Wayne, KL Harmon, and LM McIntyre. 2004. Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol Biol Evol **21**:1308-1317.

Pal, C, and LD Hurst. 2003. Evidence for co-evolution of gene order and recombination rate. Nat Genet **33**:392-395.

Poyatos, JF, and LD Hurst. 2006. Is optimal gene order impossible? Trends Genet **22**:420-423.

Richards, S, Y Liu, BR Bettencourt et al. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res **15**:1-18.

Rifkin, SA, D Houle, J Kim, and KP White. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**:220-223.

Roy, PJ, JM Stuart, J Lund, and SK Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature **418**:975-979.

Semon, M, and L Duret. 2006. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol **23**:1715-1723.

Shearwin, KE, BP Callen, and JB Egan. 2005. Transcriptional interference--a crash course. Trends Genet **21**:339-345.

Singer, GA, AT Lloyd, LB Huminiecki, and KH Wolfe. 2005. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. Mol Biol Evol **22**:767-775.

Spellman, PT, and GM Rubin. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol **1**:5.

Springer, MS, WJ Murphy, E Eizirik, and SJ O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A **100**:1056-1061.

Sproul, D, N Gilbert, and WA Bickmore. 2005. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet **6**:775-781.

Su, AI, T Wiltshire, S Batalov et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Teichmann, SA, and RA Veitia. 2004. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective. Genetics **167**:2121-2125.

Versteeg, R, BD van Schaik, MF van Batenburg, M Roos, R Monajemi, H Caron, HJ Bussemaker, and AH van Kampen. 2003. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res **13**:1998-2004.

Wang, PJ, JR McCarrey, F Yang, and DC Page. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat Genet **27**:422-426.

Waterston, RHK Lindblad-TohE Birney et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Whitehead, A, and DL Crawford. 2006. Neutral and adaptive variation in gene expression. Proc Natl Acad Sci U S A **103**:5425-5430.

Xing, Y, Z Ouyang, K Kapur, MP Scott, and WH Wong. 2007. Assessing the conservation of Mammalian gene expression using high-density exon arrays. Mol Biol Evol **24**:1283-1285.

Yanai, I, D Graur, and R Ophir. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS **8**:15-24.

Yanai, I, JO Korbel, S Boue, SK McWeeney, P Bork, and MJ Lercher. 2006. Similar gene expression profiles do not imply similar tissue functions. Trends Genet **22**:132-138.

CHAPTER 5

EFFECT OF GENE EXPRESSION EVOLUTION ON THE EVOLUTION OF NULL MUTATION PHENOTYPES

5.1 ABSTRACT

One-to-one orthologous genes of relatively closely related species are widely assumed to have similar functions and cause similar phenotypes when deleted from the genome. Although this assumption is the foundation of comparative genomics and the basis of using model organisms to study human biology and disease, its validity is known only from anecdotes rather than from systematic examination. Comparing documented phenotypes of null mutations in humans and mice, we find that over 20% of human essential genes have nonessential mouse orthologs. These changes of gene essentiality appear to be associated with adaptive evolution at the protein-sequence level while gene expression evolution plays a negligible role in such changes. Proteins localized to the vacuole, a cellular compartment for waste management, are highly enriched among essentiality-changing genes. It is probable that the evolution of the prolonged life history in humans required enhanced waste management for proper cellular function until the time of reproduction, which rendered these vacuole proteins essential and generated selective pressures on the coding sequence for their improvement. If our gene sample represents the entire genome, our results would mean frequent changes of phenotypic effects of one-to-one orthologous genes even between relatively closely related species, a possibility that should be considered in comparative

genomic studies as well as in making cross-species inferences of gene function and phenotypic effect.

5.2 INTRODUCTION

When a species diverges into two separate species, the divergent copies of a single gene in the resulting species are said to be orthologous (Fitch 1970; Koonin 2005). Although genome-wide patterns of conservation between orthologous genes have been extensively studied at the DNA and protein sequence levels (Li 1997; Nei and Kumar 2000; Koonin and Galperin 2003) and have started to be investigated at the gene expression level (Ranz et al. 2003; Jordan, Marino-Ramirez, and Koonin 2005; Khaitovich et al. 2006; Liao and Zhang 2006a), little is known about the evolutionary conservation at the levels of gene function and phenotypic effect upon gene deletion. This lack of knowledge is in part due to the widely held presumption that orthologous genes from different species are similar in function and phenotypic effect (Koonin 2005), which probably originated from a few reports that orthologous genes from distantly related species can be swapped without causing apparent phenotypic defects (Quiring et al. 1994; Lutz et al. 1996; Acampora et al. 1998; Nagao et al. 1998). Because this presumption is fundamental to comparative genomics (Koonin and Galperin 2003; Mushegian 2007) and is the basis for using model organisms such as mice to study human biology and disease (Fox 1986; Austin et al. 2004), it deserves a systematic verification.

Two model organisms, the bacterium *Escherichia coli* (Baba et al. 2006) and the yeast *Saccharomyces cerevisiae* (Winzeler et al. 1999), have been subject to genome-wide gene deletion experiments with available information on the fitness of each gene-deletion

strain, and thus could be compared in terms of the phenotypes of orthologous deletions at the genomic scale. However, these two organisms belong to prokaryotes and eukaryotes, respectively, and are so different even in basic cellular processes that the comparison is neither feasible nor meaningful. We thus choose to compare human (*Homo sapiens*) and mouse (*Mus musculus*), which are both placental mammals and have overall similar biology. Our comparison also has practical value due to the common use of mouse as a model organism for studying human biology and disease. In fact, to facilitate the use of mouse models in human biomedical research, the international genetics community recently initiated the Knockout Mouse Project (KOMP) to individually knockout every gene in the mouse genome and acquire phenotypic data (Austin et al. 2004). Our analysis will be valuable in guiding the proper use of the KOMP data.

In the present study, we focus on one of the most dramatic types of change in a gene's phenotypic effect, namely, a change in gene essentiality. A gene is said to be essential to an organism if the loss of its function renders the fitness of the organism zero; otherwise, the gene is said to be nonessential. We show that over 20% of human essential genes have nonessential mouse orthologs and elucidate the mechanisms underlying the changes of gene essentiality in evolution.

5.3 MATERIALS AND METHODS

5.3.1 Genomic data and annotations

Human genome version NCBI36 and mouse genome version NCBIM36 were used. Annotations of 31,545 human and 28,390 mouse known or predicted genes by Ensembl

(release 44) (http://www.ensembl.org/) were retrieved through BioMart (http://www.biomart.org/). We considered 14,423 pairs of human-mouse orthologous genes that were annotated as "ortholog_one2one". This annotation was not based on reciprocal best BLAST hits, but was based on phylogenetic analysis (http://www.ensembl.org/info/about/docs/compara/homology_method.html). The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse orthologs, estimated by the likelihood method, were retrieved from BioMart. The paralog information, including percent sequence identity, was also obtained from BioMart. Because retroduplicates are expected to have unrelated expression patterns from their mother genes and thus are not expected to compensate the loss of the mother genes, we did not consider retroduplicate copies as paralogs of a gene. Retroduplicates were recognized by the absence of introns that are present in their mother genes. Our results remained unchanged when retroduplicates were not excluded.

5.3.2 Human essential genes

1,716 human genes in Ensembl are associated with heritable human diseases in OMIM (http://www.ncbi.nlm.nih.gov/omim/). Among them, 1,450 have unambiguous mouse orthologs and 756 have phenotypic descriptions from gene knockout mice. Following Jimenez-Sanchez and colleagues (Jimenez-Sanchez, Childs, and Valle 2001), we categorized essentiality of a gene by the most life-threatening disease that the gene is associated with. Of the 162 human essential genes that have corresponding mouse knockout phenotypes, 24 are immunity related (MP:0005387 in MGI) and were excluded in further analysis, as the

sterilized laboratory environment may underestimate the fitness reduction associated with the deletion of immunity genes in mice. To compare with mouse knockout phenotypes, we require that human diseases considered here are due to null (or at least not gain-of-function) mutations. Null mutations are defined as nonsense or frameshift mutations or the absence of gene products in patients as determined by biochemical assays. Eighteen genes were removed because of the lack of evidence for the association between human diseases and null mutations. We manually verified the human and mouse phenotypes for the remaining 120 orthologous genes by reading relevant literature, especially ensuring that the mouse abnormal reproductive system phenotypes annotated in MGI are infertility.

5.3.3 Mouse essential and nonessential genes

Mouse phenotypic data were downloaded from MGI version 3.53 (http://www.informatics.jax.org/). We limited our analysis to null mutants generated by random gene disruption, gene trap mutagenesis, and targeted deletion, together referred to as gene knockout here. Only those genes with one-to-one matches between MGI symbol names and Ensembl gene IDs were kept for subsequent analysis. Genes with phenotypes of embryonic lethality (MP: 0002080), prenatal lethality (MP: 0002081), survival post-natal lethality (MP: 0002082), abnormal reproductive system morphology (MP: 0002160), or abnormal reproductive system physiology (MP: 0001919) were grouped as essential genes. Genes with phenotypes of premature death or induced morbidity (MP: 0002083) were manually inspected and classified according to literature. If the mutant has a lifespan of shorter than 50 days, the gene is considered to be essential. Genes associated with all other

phenotypes (at least MP: 0000001), including the normal phenotype, were grouped as nonessential genes. The dataset included 2,022 essential and 1,655 nonessential mouse genes.

5.3.4 Estimating branch-specific d_N/d_S values

Coding sequences (CDS) of H_eM_n genes from human and mouse and their "one2one" orthologs from chimpanzee (*Pan troglodytes*), macaque (*Macaca mulatta*), rat (*Rattus norvegicus*), cow (*Bos taurus*), and dog (*Canis familiaris*) were obtained from BioMart and NCBI (http://www.ncbi.nlm.nih.gov/). If multiple transcripts were annotated for one gene, the longest CDS was chosen. The sequences were aligned by MEGA4 (Tamura et al. 2007) with manual adjustment. Alignment gap sites were subsequently removed. With the known phylogeny of the seven mammals (Figure 5.2) (Murphy, Pevzner, and O'Brien 2004), the program "codeml" in PAML (Yang 2007) was used to estimate d_N/d_S for the H_eM_n orthologs in each branch of the tree, with the option of "model=1" chosen in the control file. We then compared the d_N/d_S values of the five branches (a to e in Figure5.2) that connect human and mouse in the tree. Rodents are known to have intrinsically low d_N/d_S compared to primates (Gibbs et al. 2007). To make a fair comparison of d_N/d_S among branches, we multiplied the d_N/d_S estimates for branches d and e by 1.23, which is the mean d_N/d_S value for 5,286 primate genes relative to that for their one-to-one rodent orthologs (1.28/1.04, see Figure S6 of ref. (Gibbs et al. 2007)).

5.3.5 Microarray data analysis

The GeneAtlas v2 dataset (http://symatlas.gnf.org/) contains the expression data obtained by hybridization of RNAs from 73 human non-pathogenic tissues and 61 mouse

tissues onto the Affymetrix microarray chips (human: U133A/GNF1H; mouse: GNF1M) (Su et al. 2004). We assigned the probe sets to the human and mouse genes following a previous study (Liao and Zhang 2006b). The expression level detected by each probe set was obtained as the signal intensity (*S*) computed from the MAS 5.0 algorithm. The dataset contains 26 common tissues between the two species. They are adipocyte, adrenal gland, amygdala, bone marrow, cerebellum, dorsal root ganglion, heart, hypothalamus, kidney, liver, lung, lymph node, ovary, pancreas, pituitary, placenta, prostate, skeletal muscle, spinal cord, testis, thymus, thyroid, tongue, trachea, trigeminal ganglion, and uterus. Mouse lower spinal cord was used as the homologous tissue of human spinal cord. We measured the expression-profile divergence between a pair of orthologs by 1- *R*, where *R* is Pearson's correlation coefficient between human *S* and mouse *S* across the 26 common tissues. We also used another parameter, 1-*ICE*, to measure the expression-profile divergence between orthologous genes. *ICE* (index of co-expression) between two genes is defined as the number of tissues in which both genes are expressed divided by the geometric mean of the number of tissues where each gene is expressed (Lercher, Urrutia, and Hurst 2002). Following convention (Su et al. 2004), we used a cutoff of $S = 200$ to determine whether a gene is expressed in a tissue or not when estimating 1-*ICE*.

The ExonArray data were generated using a microarray platform with over six million probes targeting all annotated and predicted exons in a genome and were obtained from Xing et al. (Xing et al. 2007). The data include six common tissues between human and mouse (heart, kidney, liver, muscle, spleen, and testis). We used the expression signals *S* computed from GeneBASE (http://biogibbs.stanford.edu/~kkapur/genebase/). The *S* values were averaged among the three replicated experiments performed for each tissue. Because

the quality of the ExonArray data is higher than that of GeneAtlas (Xing et al. 2007), a cutoff of S =150 was used to determine whether a gene is expressed in a tissue or not in the estimation of 1-*ICE*.

5.4 RESULTS AND DISCUSSION

5.4.1 Many human essential genes have nonessential mouse orthologs

From Online *Mendelian Inheritance in Man* (OMIM) (McKusick 1998), we identified 1,716 human genes with clear gene-disease associations, in which 1,450 genes have unambiguous one-to-one orthologs in the mouse genome (see Methods). This set contains 756 human genes whose mouse orthologs have been experimentally deleted with the resulting phenotypes cataloged in the database of Mouse Genome Informatics (MGI). For the 594 human genes associated with mild diseases, we cannot infer gene essentiality, because mild diseases may be due to mild mutations in essential genes or null mutations in nonessential genes. From the remaining 162 potentially essential genes, we removed 24 immunity-related genes, because the essentiality of their mouse orthologs may not have been adequately assessed in lab. We further removed 18 genes for which there is no evidence that the human disease is due to null mutations. We thus focused on the remaining 120 human genes with clinical features of death before puberty (Jimenez-Sanchez, Childs, and Valle 2001) or infertility when null mutations occur, and considered them to be essential in human. We determined the essentiality of the mouse orthologs of these human genes, based on MGI and relevant literature. Specifically, a mouse gene is considered essential if the knockout mice cannot survive to reproductive age (50 days) or are infertile.

To our surprise, 27 (22.5%) of the 120 mouse orthologs of human essential genes are nonessential (Table 5.1). Furthermore, except for reduced survival or fecundity for *Mthfr*, *Smpd1 Hexb*, and *Neu1*-knockout mice, the other 23 knockout mouse strains (19.2%) are able to breed as successfully as the wild-type at least up to the age of 6 months (Table 1). For convenience, we term these 27 human-essential-mouse-nonessential orthologs as H_eM_n orthologs and the other 93 human-essential-mouse-essential orthologs as H_eM_e orthologs, where H and M indicate human and mouse, respectively, and the subscripted "e" and "n" indicate essential and nonessential genes, respectively.

5.4.2 Gene duplication is not the cause of gene-essentiality changes

What caused the dramatic change in essentiality between human and mouse in over 20% of the examined genes? Previous studies in yeast (Gu et al. 2003) and nematode (Conant and Wagner 2004) suggested that when a gene is deleted, its paralogous gene(s) can often provide functional compensation such that an otherwise essential gene would appear to be nonessential. The H_eM_n and H_eM_e orthologs studied here are one-to-one orthologs and hence do not have paralogs that were generated since the human-mouse separation. Nevertheless, it is possible that a paralog that was generated before the human-mouse separation is retained in mouse but lost in human, rendering the effect of functional compensation present in mouse but absent in human. We thus examined whether H_eM_n-type mouse genes tend to have (i) more paralogs and (ii) closer paralogs in the mouse genome than H_eM_e-type mouse genes, which could explain why some orthologs of human essential genes are nonessential in mouse. We found that the proportion of mouse genes that have paralogs is not significantly different between the H_eM_n group (18/27=66.7%) and the H_eM_e

group (55/93=59.1%) (P=0.512, Fisher's exact test). Among the mouse genes that have paralogs, H_eM_n-type mouse genes do not have significantly more paralogs (average number of paralogs = 4.33) than H_eM_e-type mouse genes have (average = 3.78) (P=0.415, Mann-Whitney U test). Moreover, H_eM_n-type mouse genes are not more similar to their closest paralogs (average protein sequence identity = 56.2%) than H_eM_e-type mouse genes are to their closest paralogs (average = 58.3%; P=0.568, U test). Because divergent paralogs are unlikely to compensate one another, we repeated our analysis by considering only paralogs with relatively high protein sequence identities, but our results remain unchanged (Table A.6). These observations, consistent with recent reports of a general lack of functional compensation between paralogs in mammals (Liang and Li 2007; Liao and Zhang 2007), indicate that the dramatic changes of gene essentiality between human and mouse orthologs are not due to differential functional compensation from paralogs. Rather, it is more likely that the evolutionary changes of gene essentiality have resulted from alterations of the genes themselves.

5.4.3 Gene essentiality changes are associated with accelerated protein sequence evolution

Were changes of gene essentiality more frequently caused by alterations of protein function or gene expression? To address this question, we first estimated the nonsynonymous distance (d_N) between each pair of human and mouse orthologs. We found that d_N of the H_eM_n group is significantly greater than that of the H_eM_e group ($P = 5.04\times10^{-5}$, U test) (Figure 5.1a), whereas the synonymous distances (d_S) are not significantly different between the two groups (P = 0.697, U test) (Figure 5.1b). For comparison, let us also define

an H_aM_n group, which includes any human-mouse orthologous pair in which the mouse gene is known to be nonessential. Our H_aM_n group comprises 864 nonessential non-immune-system mouse genes and their human orthologs. The relatively large d_N of the H_eM_n group compared to that of the H_eM_e group must be due to accelerated nonsynonymous substitutions caused by (i) weaker purifying selection in the mouse lineage on H_eM_n genes than on H_eM_e genes, because mammalian nonessential genes are subject to weaker purifying selection and consequently have higher d_N than essential genes (Liao, Scott, and Zhang 2006), and/or (ii) positive selection associated with the change of function and essentiality of H_eM_n genes. If (i) is the primary reason, the d_N of the H_eM_n group should be lower than that of the H_aM_n group, because the latter is composed of H_nM_n and H_eM_n genes. However, we found that the d_N of H_eM_n genes is significantly greater than that of H_aM_n genes ($P = 1.62 \times 10^{-4}$, *U* test; Figure 5.1a), suggesting that (i) cannot be the primary reason of greater d_N for H_eM_n genes than for H_eM_e genes. Consequently, (ii) must have contributed to a large degree. As expected, there is no significant difference in d_S between H_eM_n and H_aM_n genes (P =0.770, *U* test) (Figure 5.1b).

To reconfirm (ii), we used a maximum-likelihood method (Yang 1998) to estimate branch-specific d_N/d_S values in a phylogeny of seven placental mammals for each of the 27 H_eM_n genes (Figure 5.2). Besides human and mouse, five additional mammals (chimpanzee, macaque, rat, dog, and cow) were chosen because they can divide the evolutionary path linking human and mouse and because they have publicly available high-quality (i.e., at least 6× coverage) genome sequences so that the orthologous sequences of the H_eM_n genes can be retrieved. We then compared the d_N/d_S values for the five branches connecting human and mouse (Figure 5.2). An earlier genomic study showed that orthologous genes have on

average lower d_N/d_S in rodents than in primates, likely due to a larger population size and consequently increased efficacy of purifying selection in rodents than in primates (Gibbs et al. 2007). To make a fair comparison here, we multiplied the estimated d_N/d_S values for the two rodent branches (d and e in Figure 5.2) by 1.23, which is the mean d_N/d_S value for 5,286 primate genes relative to that for their one-to-one rodent orthologs analyzed in an earlier study (Gibbs et al. 2007). Under no positive selection, H_eM_n genes are expected to exhibit relatively low d_N/d_S values in branches closer to human and relatively high d_N/d_S values in branches closer to mouse, along the evolutionary path connecting human and mouse, because essential genes tend to have lower d_N/d_S than nonessential genes (Liao, Scott, and Zhang 2006). We, however, observed the opposite pattern. That is, the fraction of H_eM_n genes that have their highest d_N/d_S values in the two branches closest to human (a and b in Figure 5.2) is significantly greater than the chance expectation of 2/5 (P=0.014, binomial test). A recent analysis of a high-exchangeability group of amino acid changes suggests that $d_N/d_S > 0.5$ likely indicates positive selection (Tang and Wu 2006). Again, we found that the fraction of incidences where d_N/d_S of an H_eM_n gene is larger than 0.5 in branch a or b is greater than the chance expectation (P=0.004, binomial test; Figure 5.2). The same is true when both highest d_N/d_S in a branch and d_N/d_S >0.5 are considered (P=0.015, binomial test; Figure 5.2). Taken together, these results suggest that accelerated protein sequence evolution driven by positive selection was associated with changes of gene essentiality in at least an appreciable fraction of H_eM_n genes and that most H_eM_n genes had their gene essentiality changed during primate evolution.

5.4.4 Gene expression evolution is not the cause of gene essentiality changes

Next, we measured the expression-profile divergence between human and mouse orthologous genes by 1-*R*, where *R* is Pearson's correlation coefficient between their expression levels across homologous tissues of the two species (see Methods). Two independent microarray gene expression datasets were used. The ExonArray dataset has a higher accuracy in interspecific comparisons (Xing et al. 2007), while the GeneAtlas v2 dataset contains more homologous tissues between the two species (Su et al. 2004). Neither dataset shows a significant difference in 1-*R* between H_eM_n genes and H_eM_e genes (*P* = 0.230 and 0.140 in Figure 5.1c and 5.1d, respectively, *U* test). Furthermore, 1-*R* is not significantly different between H_eM_n and H_aM_n genes in these datasets (*P* = 0.433 and 0.420 in Figure 5.1c and 1d, respectively, *U* test). In short, we did not find accelerated gene expression evolution to be associated with the essentiality changes of H_eM_n genes. Use of other measures of gene expression divergence gave similar results (Figure A.16).

5.4.5 Gene essentiality changes and the vacuole

To better understand the biological reasons behind the changes of gene essentiality, we compared the Gene Ontology of the human genes in the H_eM_n group and the H_eM_e group using FatiGO (Al-Shahrour, Diaz-Uriarte, and Dopazo 2004). There is only one category that is significantly different between the two groups after the control for multiple testing. A much greater fraction of H_eM_n genes (12/27 = 44.4%) than H_eM_e genes (4/93 = 4.3%) have their protein products localized to the vacuole (false discovery rate $q = 5.52\times10^{-5}$), a cellular compartment primarily responsible for containing and degrading wastes and toxins. The absence of vacuole proteins in humans tends to cause the accumulation of cellular wastes and toxins that often leads to fatal neurological diseases (Table 1). The mass-corrected basal

metabolic rate in human is ~12% of that in mouse (Tolmasoff, Ono, and Cutler 1980), but human reproductive age is ~150 times that of mouse (Table A.7). Consequently, the total amount of waste produced till reproduction for every gram of body mass is ~18 times higher for human than for mouse. Hence, waste management is much more important in human than in mouse for maintaining proper cellular functions until the time of reproduction. This may have rendered the orthologs of many nonessential mouse vacuole proteins essential in humans. Consistent with this idea, deficiencies of vacuole proteins tend to cause defects at a later life stage in mouse than in human (Table 1). Furthermore, the evolution of the prolonged life history of humans probably generated selective pressures for better vacuole proteins, which may be part of the reason behind the accelerated protein sequence evolution observed in H_eM_n genes. Comparison of the product of the metabolic rate and the starting reproductive age among primates suggests that the importance of vacuole functions gradually increased in the primate lineage leading to humans, beginning from the common ancestor of all extant primates (Table A.7). Consistent with this pattern, H_eM_n vacuole proteins show accelerated sequence evolution in the three primate branches (a, b, and c) in Figure 5.2. However, due to the small sample size, only one comparison yielded statistically significant enrichment in the three branches. That is, for H_eM_n vacuole proteins, incidences of branch-specific d_N/d_S >0.5 occurs more frequently in these three branches than expected by chance (P=0.028, binomial test).

About 55% of H_eM_n genes are not vacuole proteins. We confirmed that the results in Figure 5.1 remain qualitatively unchanged after the removal of vacuole proteins (Figure A.17). Although the biological reason behind the change of gene essentiality of these non-vacuole proteins is unclear, the association between the essentiality change and accelerated

protein sequence evolution may be similarly caused by an increase in the importance of a particular biological process during human evolution since the human-mouse split, which rendered a nonessential gene essential and at the same time generated selective pressures for the improvement of the gene function. Consistent with this idea, analysis of branch-specific d_N/d_S indicates that non-vacuole H_eM_n proteins are significantly more likely to have rapid evolution and highest d_N/d_S in branch a or b of Figure 5.2 than expected by chance (P=0.002, 0.019, and 0.019, respectively, for the three properties shown in Figure 5.2, binomial test).

5.4.6 Final remarks

It is possible that the frequency and direction of gene essentiality changes are not the same among evolutionary lineages. For example, the proportion of essential genes in a genome is much greater in mouse than in yeast, which is in turn much greater than that in *E. coli* (Liao and Zhang 2007). In the present work, although only H_eM_n genes are systematically examined, anecdotes of H_nM_e genes are known. For example, humans with homozygous *RECQL* null alleles display viable and fertile Bloom's syndrome, while targeted deletion of the ortholog in mouse causes embryonic lethality (Chester et al. 1998). Unfortunately, it is not possible to identify H_nM_e genes systematically, owing to the difficulty in proving the non-essentiality of human genes. This obstacle notwithstanding, it is almost certain that the prevalence of distinct null phenotypes of human and mouse orthologs is underestimated here. The first reason is that genes with unaltered essentiality could still have altered phenotypic effects. For instance, *Adamts2*, *Acox1*, and *Fancg* are considered essential for human due to the mutant phenotype of premature death (Jimenez-Sanchez, Childs, and Valle 2001; Suzuki et al. 2002), but they are essential for mouse due to the knockout

phenotype of infertility of adult mice (Fan et al. 1996; Li et al. 2001; Yang et al. 2001). Second, the phenotypes associated with nonessential genes are probably more labile in evolution than those associated with essential genes, because changes of nonessential genes are expected to be more tolerable than changes of essential genes. Therefore, it is likely that significantly more than 20% of one-to-one orthologs between human and mouse have different phenotypic effects when deleted. However, we caution that the gene sample analyzed here is relatively small and thus our results should be reconfirmed when more data become available. In the future, it may also be possible to verify our results by comparing the essentiality of one-to-one orthologous genes from several bacterial species that have been subject to genome-wide gene deletion experiments (Akerley et al. 2002; Baba et al. 2006; Gallagher et al. 2007). However, due to high incidences of horizontal gene transfer (Doolittle 1999) and non-orthologous gene replacement (Koonin, Mushegian, and Bork 1996) in prokaryotes, caution should be taken in such comparisons. When studying functional changes in orthologous gene evolution, it is important to distinguish among changes of molecular function, changes of involved biological processes, and changes of physiological importance. By comparing gene essentiality, we are addressing the physiological importance of a gene. A careful examination of Table 1 suggests that the molecular functions and the involved biological processes are likely to be unaltered for the majority of the 27 H_eM_n genes, while their physiological importance has changed dramatically.

Potential implications of our findings are manifold. First, gene annotation based on mutant phenotypes in other species may often be wrong, especially about gene essentiality. Second, comparative and evolutionary analysis dependent on the assumption of conservation of gene function or importance between orthologs should be interpreted carefully. Third,

alteration of gene essentiality between species could be a cause of the observation that some mutations pathogenic to one species are nevertheless fixed in other species (Kondrashov, Sunyaev, and Kondrashov 2002; Gao and Zhang 2003). Fourth, it is possible that mouse models of a large number of human diseases will not yield sufficiently accurate information, although they might provide some basic knowledge. The scientific community may need to strategically and systematically consider establishing a primate model organism for studying many human diseases. In this regard, it is particularly important to choose appropriate animal models for the study of human neurological disorders that involve malfunctioning vacuole proteins, due to the opposite essentiality of many vacuole proteins between human and mouse. Finally, the association between changes of gene essentiality and the prolonged life history of humans sheds light on the mechanisms of some human-specific disorders that accompany apparently beneficial human traits.

Although a recent literature survey found otherwise (Hoekstra and Coyne 2007), many believe that changes of gene expression are more important than changes of protein function in generating phenotypic differences between species (King and Wilson 1975; Carroll 2005). We found that changes of gene essentiality were accompanied by accelerated evolution that was likely driven by positive selection at the protein sequence level, but did not find such a signal at the gene expression level. Although we cannot exclude the possibility that our result regarding expression evolution is caused by the relatively large noise of microarray expression data or the lack of relevant tissues in the datasets analyzed, we can conclude that protein sequence and function changes are important in the change of gene essentiality in evolution. It remains possible, however, that gene expression changes

are more important for phenotypic evolution that does not involve a change in gene essentiality.

5.5 ACKNOWLEDGMENTS

We thank Gerardo Jimenez-Sanchez for providing details of the annotation of human genetic diseases. Meg Bakewell, Wendy Grus, David Lipman, Wenfeng Qian, and three anonymous referees made valuable comments. This work was supported by a pilot grant from University of Michigan Center for Computational Medicine and Biology and research grants from National Institutes of Health to J.Z. B.Y.L was supported by the University of Michigan Rackham Predoctoral Fellowship.

Figure 5.1 Sequence divergence and expression divergence of human-mouse orthologs. The quartile-plots of sequence divergence (a: d_N, b: d_S) and expression-profile divergence (c: ExonArray, d: GeneAtlas v2) between human and mouse one-to-one orthologous genes. Values of upper quartile, median, and lower quartile are indicated in each box. The bars indicate semi-quartile ranges. H and M indicate human and mouse, respectively, and the subscripts e, n, and a indicate essential, nonessential, and any genes, respectively. The *P*-vales are determined by two-tail Mann-Whitney *U* tests.

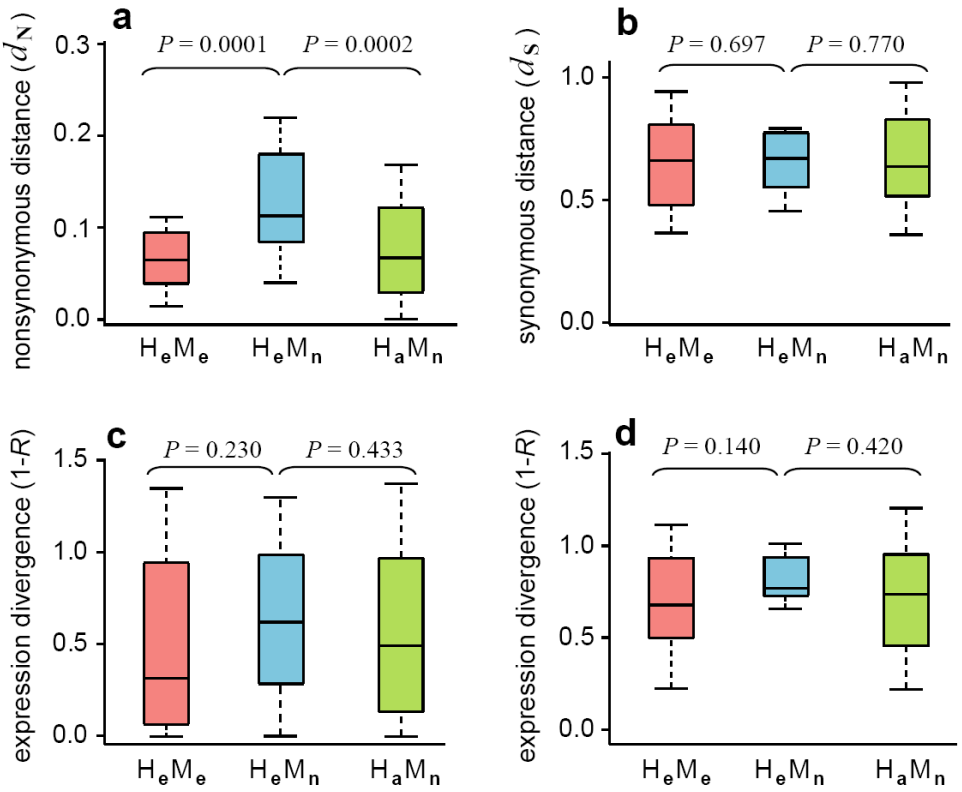
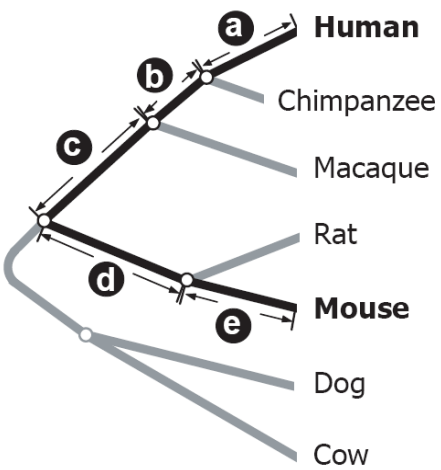


Figure 5.2 Natural selection one the branches leading to human and mouse. Variation of branch-specific d_N/d_S values among the five branches (marked a to e) that connect human and mouse in the mammalian phylogeny. The branch lengths are not drawn to scale. The d_N/d_S values for branches d and e have been adjusted to correct for the intrinsically low d_N/d_S in rodents (see Methods).



Numbers of H_eM_n genes (vacuole + non-vacuole) with

	Highest d_N/d_S	d_N/d_S > 0.5	Highest d_N/d_S & d_N/d_S > 0.5
a	8 (2+6)	6 (3+3)	5 (2+3)
b	9 (3+6)	5 (2+3)	4 (1+3)
c	3 (2+1)	2 (2+0)	2 (2+0)
d	2 (2+0)	0 (0+0)	0 (0+0)
e	5 (3+2)	1 (0+1)	1 (0+1)

Table 5.1. One-to-one orthologous genes that are essential in human but nonessential in mouse

Human gene name	Human disease name	Mouse gene knock-out phenotypes
Arylsulfatase A (ARSA) [a]	Metachromatic leukodystrophy [b]	normal fertility and litter size; impaired balance and spatial learning ability; sulfatide accumulation in the white matter of the brain; reduced myelin sheath thickness in the corpus callosum & optic nerves; a low frequency head tremor develops > 2 years of age
Alpha-mannosidase , class 2B, member 1 (MAN2B1)	Mannosidosis, alpha-, types I and II [b]	normal development , fertility; no elevated mortality; mild form of human alpha-mannosidosis
Dystrophia myotonica protein kinase (DMPK) [a]	Myotonic dystrophy-1 [b]	normal fertility and litter size; abnormal sodium channel gating in cardiac myocytes; cardiac conduction defects; late-onset progressive skeletal myopathy
Lysosomal acid lipase (LIPA)	Wolman disease	normal development and fertility; accumulation of triglycerides and cholesteryl esters occurs in several organs
Axonemal heavy chain dynein type 11 (DNAH11)	Primary ciliary dyskinesia; Kartagener syndrome [c]	normal fertility; abnormal left-right axis patterning
Sialyltransferase 9 (ST3GAL5)	Amish infantile epilepsy syndrome [b]	normal viability and fertility; hypoglycemia; increased insulin sensitivity; abnormal lipid level
Patched homolog 2 (PTCH2)	Medulloblastoma; Basal cell carcinoma [b]	normal viability and fertility; normal cell proliferation or differentiation in the cerebellum; abnormal dermal morphology in some males
Granulocyte colony-stimulating factor (CSF3R)	Kostmann neutropenia [b]	normal development and fertility; reduced numbers of peripheral neutrophils; fewer hematopoietic progenitors in bone marrow
5,10-methylenetetrahydrofolate reductase (MTHFR)	Homocystinuria due to MTHFR deficiency [b]	reduced survival rate but fertile; delayed development; elevated plasma levels of homocysteine
Transforming growth factor-beta interacting factor (TGIF1)	Holoprosencephaly-4 [b]	normal growth, behavior and fertility
Aacid phosphatase-2 (ACP2) [a]	Acid phosphatase deficiency [b]	normal development and fertility; skeletal defects in mutants greater then 6 months of age; a small percentage of mutants exhibit tonic-clonic seizures
Cathepsin A (CTSA) [a]	Galactosialidosis [b]	normal fertility; death occurs at ~12 months; aberrant lysosomal storage; enlarged spleen and liver; abormally flat face; reduced body size; generalized edema, ataxia and tremors
N-acetylglucosaminidase (NAGLU) [a]	Sanfilippo syndrome, type B [b]	appear normal, healthy, and fertile up to 6 months of age; survive to 8-12 months; reduced open field activity; massive accumulation of heparan sulfate in kidney and liver; elevated gangliosides in brain; vacuoles in macrophages, epithelial cells, and neurons.
Beta-mannosidase (MANBA) [a]	Beta-Mannosidosis [b]	normal appearance, growth, and fertility to 1 year of age; cytoplasmic vacuolation in central nervous system and visceral organs
Ubiquitin-protein ligase e3 component n-recognin 1 (UBR1)	Johanson-Blizzard syndrome [b]	normal viability and fertility; 20% lower body weight; reduced muscle and adipose tissue; abnormal metabolism; enhanced non-spatial learning; impaired spatial learning
von Willebrand factor-cleaving protease (ADAMTS13)	Congenital thrombotic thrombo-cytopenic purpura [b]	normal development, viability, and fertility; prolonged vWF-mediated platelet-endothelial interactions
Acid sphingomyelinase (SMPD1) [a]	Niemann-Pick disease, type A and B [b]	males could breed until 20 weeks of age and females until 10 weeks of age with normal litter size; lifespan of 4-8 months; impaired coordination; decreased body weight
Beta-glucosidase-1 (GLB1) [a]	GM1-gangliosidosis; Mucopolysaccharidosis IVB [b]	normal fertility and litter size; lifespan of 7-10 months; progressive spastic diplegia; emaciation; accumulation of ganglioside GM1 and asialo GM1 in brain tissue

Table 5.1. (continued)

Human gene name	Human disease name	Mouse gene knock-out phenotypes
Alpha-1,4-glucosidase (GAA) [a]	Glycogen storage disease II [b]	normal growth and fertility; reduced mobility and strength; impaired coordination, hindlimb paralysis and muscle weakness for the mutants older than 8 months of age
Cytochrome p450, family 7, subfamily b, polypeptide 1 (CYP7B1) [a]	Giant cell hepatitis, neonatal [b]	normal survival, physical appearances, and behaviors; normal bile acid metabolism, plasma cholesterol and triglyceride levels; sterol biosynthetic rates were unaffected in multiple tissues with the exception of the male kidney, which showed a ~40% decrease
Coagulation factor VIII (F8)	Hemophilia A [b]	females exhibit normal fertility and pregnancy; males show reduced ability to clot blood; no spontaneous bleeding into joints or soft tissues is observed up to 12 weeks of age
Hexosaminidase B (HEXB)	Sandhoff disease [b]	normal growth and fertility; mutants exhibit spasticity, muscle weakness, rigidity, tremors, and ataxia beginning around 4 months of age and resulting in death about 6 weeks later
GM2 activator protein (GM2A) [a]	GM2-gangliosidosis, AB variant [b]	normal growth, survival and fertility; abnormal accumulation of glycolipid and ganglioside in various brain regions with impaired balance, coordination, and learning
Very long-chain acyl-CoA dehydrogenase (ACADVL)	Deficiency of Acyl-CoA dehydrogenase, VL [b]	normal gross appearance, survival, behavior and fertility; normal body and heart weight at 2 months of age.
Alanine:glyoxylate aminotransferase (AGXT)	Hyperoxaluria, primary, type 1 [b]	normal growth and development; no histological differences between mutants and wild types in multiple tissues; increased oxalate urine levels and higher chance to develop bladder stones for males.
Neuraminidase 1 (NEU1) [a]	Sialidosis, type I and type II [b]	27% of the pups in the NMRI background and 10–15% in the C57BL/6 background died suddenly around weaning age; mice that survived past the 21 days were fertile, but stopped producing offspring by the age of 10 weeks; death occurred between the ages of 8 and 12 months.
Galactose-1-phosphate uridylyltransferase (GALT)	Galactosemia [b]	normal embryonic survival; normal fertility in both sexes; abnormal galactose metabolism, but lack symptoms of acute toxicity.

[a] Protein product localized to vacuole
[b] Death before puberty
[c] Infertility

5.6 LITERATURE CITED

Acampora, D, V Avantaggiato, F Tuorto, P Barone, H Reichert, R Finkelstein, and A Simeone. 1998. Murine Otx1 and Drosophila otd genes share conserved genetic functions required in invertebrate and vertebrate brain development. Development **125**:1691-1702.

Akerley, BJ, EJ Rubin, VL Novick, K Amaya, N Judson, and JJ Mekalanos. 2002. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A **99**:966-971.

Al-Shahrour, F, R Diaz-Uriarte, and J Dopazo. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics **20**:578-580.

Austin, CP, JF Battey, A Bradley et al. 2004. The knockout mouse project. Nat Genet **36**:921-924.

Baba, T, T Ara, M Hasegawa et al. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol **2**:2006 0008.

Carroll, SB. 2005. Evolution at two levels: on genes and form. PLoS Biol **3**:e245.

Chester, N, F Kuo, C Kozak, CD O'Hara, and P Leder. 1998. Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. Genes Dev **12**:3382-3393.

Conant, GC, and A Wagner. 2004. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. Proc Biol Sci **271**:89-96.

Doolittle, WF. 1999. Phylogenetic classification and the universal tree. Science **284**:2124-2129.

Fan, CY, J Pan, R Chu et al. 1996. Hepatocellular and hepatic peroxisomal alterations in mice with a disrupted peroxisomal fatty acyl-coenzyme A oxidase gene. J Biol Chem **271**:24698-24710.

Fitch, W. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. **19**:99-106.

Fox, M. 1986. The Case for Animal Experimentation: An Evolutionary and Ethical Perspective. University of California Press, Berkeley, CA.

Gallagher, LA, E Ramage, MA Jacobs, R Kaul, M Brittnacher, and C Manoil. 2007. A comprehensive transposon mutant library of Francisella novicida, a bioweapon surrogate. Proc Natl Acad Sci U S A **104**:1009-1014.

Gao, L, and J Zhang. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet **19**:678-681.

Gibbs, RAJ RogersMG Katze et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**:222-234.

Gu, Z, LM Steinmetz, X Gu, C Scharfe, RW Davis, and WH Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature **421**:63-66.

Hoekstra, HE, and JA Coyne. 2007. The locus of evolution: evo devo and the genetics of adaptation. Evolution Int J Org Evolution **61**:995-1016.

Jimenez-Sanchez, G, B Childs, and D Valle. 2001. Human disease genes. Nature **409**:853-855.

Jordan, IK, L Marino-Ramirez, and EV Koonin. 2005. Evolutionary significance of gene expression divergence. Gene **345**:119-126.

Khaitovich, P, W Enard, M Lachmann, and S Paabo. 2006. Evolution of primate gene expression. Nat Rev Genet **7**:693-702.

King, MC, and AC Wilson. 1975. Evolution at two levels in humans and chimpanzees. Science **188**:107-116.

Kondrashov, AS, S Sunyaev, and FA Kondrashov. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A **99**:14878-14883.

Koonin, E, and M Galperin. 2003. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics Kluwer Academic Publishers, Boston.

Koonin, EV. 2005. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet **39**:309-338.

Koonin, EV, AR Mushegian, and P Bork. 1996. Non-orthologous gene displacement. Trends Genet **12**:334-336.

Lercher, MJ, AO Urrutia, and LD Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet **31**:180-183.

Li, SW, M Arita, A Fertala, Y Bao, GC Kopen, TK Langsjo, MM Hyttinen, HJ Helminen, and DJ Prockop. 2001. Transgenic mice with inactive alleles for procollagen N-proteinase (ADAMTS-2) develop fragile skin and male sterility. Biochem J **355**:271-278.

Li, W-H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Liang, H, and WH Li. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet **23**:375-378.

Liao, BY, NM Scott, and J Zhang. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol **23**:2072-2080.

Liao, BY, and J Zhang. 2007. Mouse duplicate genes are as essential as singletons. Trends Genet **23**:378-381.

Liao, BY, and J Zhang. 2006a. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol **23**:530-540.

Liao, BY, and J Zhang. 2006b. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol **23**:1119-1128.

Lutz, B, HC Lu, G Eichele, D Miller, and TC Kaufman. 1996. Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev **10**:176-184.

McKusick, VA. 1998. Mendelian Inheritance in Man. Johns Hopkins University Press, Baltimore.

Murphy, WJ, PA Pevzner, and SJ O'Brien. 2004. Mammalian phylogenomics comes of age. Trends Genet **20**:631-639.

Mushegian, A. 2007. Foundations of Comparative Genomics. Academic Press, Burlington, Mass.

Nagao, T, S Leuzinger, D Acampora, A Simeone, R Finkelstein, H Reichert, and K Furukubo-Tokunaga. 1998. Developmental rescue of Drosophila cephalic defects by the human Otx genes. Proc Natl Acad Sci U S A **95**:3737-3742.

Nei, M, and S Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Quiring, R, U Walldorf, U Kloter, and WJ Gehring. 1994. Homology of the eyeless gene of Drosophila to the Small eye gene in mice and Aniridia in humans. Science **265**:785-789.

Ranz, JM, CI Castillo-Davis, CD Meiklejohn, and DL Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science **300**:1742-1745.

Su, AI, T Wiltshire, S Batalov et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Suzuki, Y, M Iai, A Kamei et al. 2002. Peroxisomal acyl CoA oxidase deficiency. J Pediatr **140**:128-130.

Tamura, K, J Dudley, M Nei, and S Kumar. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**:1596-1599.

Tang, H, and CI Wu. 2006. A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection. Mol Biol Evol **23**:372-379.

Tolmasoff, JM, T Ono, and RG Cutler. 1980. Superoxide dismutase: correlation with life-span and specific metabolic rate in primate species. Proc Natl Acad Sci U S A **77**:2777-2781.

Winzeler, EA, DD Shoemaker, A Astromoff et al. 1999. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science **285**:901-906.

Xing, Y, Z Ouyang, K Kapur, MP Scott, and WH Wong. 2007. Assessing the conservation of Mammalian gene expression using high-density exon arrays. Mol Biol Evol **24**:1283-1285.

Yang, Y, Y Kuang, R Montes De Oca, T Hays, L Moreau, N Lu, B Seed, and AD D'Andrea. 2001. Targeted disruption of the murine Fanconi anemia gene, Fancg/Xrcc9. Blood **98**:3435-3440.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15**:568-573.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**:1586-1591.

CONCLUSIONS

Comparing the transcriptomes of humans and mice, I address fundamental questions about mammalian gene expression. In the first section of my dissertation (Chapter 1-2), I confirm the presence of evolutionary constraint on gene expression and study the evolutionary patterns of mammalian gene expression.

In Chapter 1, contrary to what the neutral model of transcriptome evolution asserts, I demonstrate that over 80% of human-mouse orthologous genes are evolutionarily conserved in their expression-profiles, highlighting the importance of proper gene expression to fitness.

If gene expression is generally constrained by natural selection, genes with different properties may have experienced different selective constraints. In Chapter 2, I show that the expression-profiles of highly expressed and tissue-specific genes tend to evolve slowly, implying that the expression pattern is of particular importance to highly expressed and tissue-specific genes. Comparison of the rate determinants for protein evolution and that for expression-profile evolution shows that different rules apply to the evolution of protein sequences and that of expression-profiles.

Considering several technical issues related to the analysis of microarray data in cross-species comparisons, the conclusions made in the first two chapters are contradictory to those of many previous studies. The inconsistency of results between my studies and previous studies highlights the importance of choosing proper indices and developing new statistical approaches for comparative transcriptomics. My thesis specifically addresses the evolution of gene expression profiles rather than the evolution of gene expression levels.

The main reason is that, compared to sequencing-based gene expression profiling, hybridization-based methods are less reliable in measuring absolute quantities of transcripts (see Chapter 1). In the coming years, methods of sequencing-based gene expression profiling (e.g. MPSS, Massively Parallel Signature Sequencing) (Brenner et al. 2000) will mature and become more affordable for generating comprehensive datasets. At that time, mammalian gene expression can be understood from both aspects of expression level and expression profile such that a comprehensive picture of gene expression evolution may be revealed.

In the second section of my dissertation (Chapter 3-5), I investigate the potential roles gene expression plays in protein sequence evolution, dynamics of genome organization, and the evolutionary changes of gene essentiality.

Studies based on yeast species suggested that the level of gene expression is the most important factor determining the rate of protein sequence evolution. In Chapter 3, analysis of mammalian genes, however, shows that tissue-specificity, a characteristic of gene expression found only in multicellular organisms, is a stronger determinant of protein evolutionary rate. When compared to gene compactness, gene essentiality and tissue-specificity, gene expression level is the least important factor in mammals. Thus, there is a great difference in rate determinants of protein evolution between unicellular and multicellular organisms. It should be noted that although yeast proteins doe not have tissue-specificity, some proteins are heterogeneously expressed under different conditions. In the future, it would be interesting to examine whether condition-specific yeast genes, as mammalian tissue-specific genes, also evolve rapidly if other confounding factors are controlled for.

In Chapter 4, I study the impact of gene expression on genome organization. It has been reported in various eukaryotes that genes with physical proximity in a chromosome tend to have similar expression patterns. Previous authors assumed that this phenomenon is a result of adaptive relocation of initially unlinked but co-expressed genes, but this assumption is incompatible with several observations. I propose that co-expression of linked genes is a form of transcriptional interference that is disadvantageous to the organism. My hypothesis is supported by genome-wide analyses of co-expression, recombination, and chromosomal rearrangements in mammalian genomes. My model suggests that transcriptional interference is the main cause of co-expressed gene clusters and may promote recurrent relocations of genes in the genome.

In Chapter 5, I compare documented phenotypes of null mutations in humans and mice and find that over 20% of human essential genes have nonessential mouse orthologs. These changes of gene essentiality appear to be associated with adaptive evolution at the protein-sequence level, while gene duplication and gene expression evolution plays a negligible role. In light of the finding that the proteins localized to the vacuole are highly enriched among essentiality-changing genes, I hypothesize that the evolution of the prolonged life history in humans may have rendered these vacuole proteins essential and generated selective pressures on the coding sequence for their improvement.

Results from the second part of my dissertation have special implications for genomic studies using bioinformatic approaches. Chapter 3 reveals an unexpected diversity in the rules governing protein sequence evolution among different organisms. From an evolutionary perspective, any two organisms may share certain biological similarities, including rules of evolution, due to ancestry. My results suggest the necessity of confirming

the existence of such similarities between organisms, especially before one applies the empirical observations from a distantly-related model organism to the species of interest. The conclusions from Chapter 5, which investigates the phenotypes of human-mouse orthologs, further strengthen this argument.

What causes the variation in general rules governing molecular evolution between two types of organisms (e.g. unicellular vs. multicellular eukaryotes) is an important question that requires further investigation. Our results clearly indicate that lineage-specific properties can have stronger impact than ancestral properties on protein sequence evolution, and perhaps organismal evolution as well. For instance, tissue-specificity, a property present in mammals but not yeasts, was shown to be a stronger rate determinant of mammalian protein evolution than expression level (see Chapter 3). Also, the primate-specific trait of elongated life history may underlie the positive selection in primate vacuole proteins (see Chapter 5). Do the processes selecting biological novelties also change the previously established "rules" for organismal evolution? How flexible are such rules during the evolution? It would be interesting to know how lineage-specific properties emerged and how much impact they have on the molecular evolution of genomes compared to ancestrally derived shared properties. As more phylogenetically comprehensive genomic, transcriptomic and phenomic data become available, such studies can be initiated by comparing closely related species and then expanding the investigations to a phylogenetically larger scale.

The abundance of mRNA is determined by levels of transcription and mRNA degradation. MicroRNAs (miRNAs) can facilitate the degradation of targeted mRNA (Tolia and Joshua-Tor 2007). Although Chapter 4 suggests that co-expression of linked genes likely occurs at the transcriptional level, in light of recent discoveries that miRNAs are a

factor determining organismal complexity (Heimberg et al. 2008), the important role of miRNAs in the evolution of gene expression cannot be neglected. In the future, mechanisms driving the divergence of gene expression can be studied when additional empirical data on mRNA production and degradation become available. Because most nucleotides in mammalian genomes are non-coding (Venter et al. 2001; Waterston et al. 2002) and potentially contain the information required for gene regulation, there is a pressing need for understanding the molecular function and evolutionary dynamics of non-coding sequences.

LITERATURE CITED

Brenner S, Johnson M, Bridgham J, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol **18**:630-634.

Heimberg AM, Sempere LF, Moy VN, et al. 2008. MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci U S A **105**:2946-2950.

Tolia NH, and Joshua-Tor L. 2007. Slicer and the argonautes. Nat Chem Biol **3**:36-43

Venter JC, Adams MD, Myers EW, et al. 2001. The sequence of the human genome. Science **291**:1304-1351.

Waterston RH, Lindblad-Toh K, Birney E, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

APPENDIX

SUPPLEMENTARY FIGURES AND TABLES

Figure A.1 Pairwise comparison of expression profiles of two probe sets of the same human genes. The 3,762 genes in group A contain both optimal and suboptimal probe sets, whereas the 1,097 genes in group B contain only optimal probe sets. The distributions show that the expression profiles detected by two probe sets of the same gene are more similar for group A genes than for group B genes ($P<10^{-27}$, Mann-Whitney U test), implying that "suboptimal" probe sets produce more consistent expressional profiles than "optimal" probe sets.

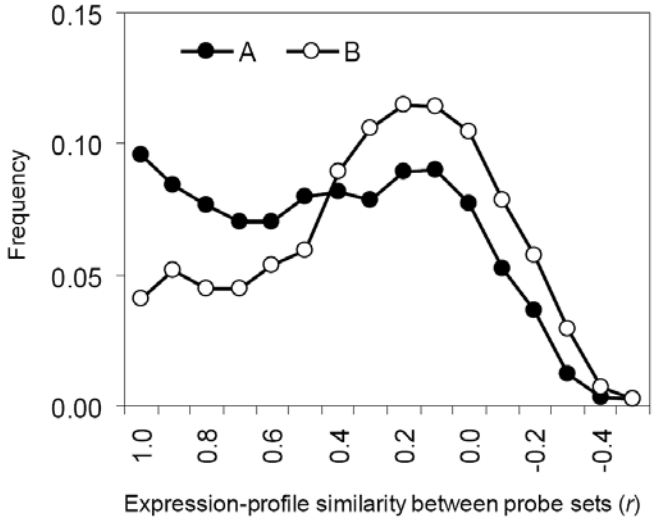


Figure A.2 Net distances (*D*) of expressional profiles between human and mouse orthologs and Euclidean distances (*d*) of random human-mouse gene pairs. The distribution of the random pairs represents the neutral expectation of expressional divergences. The black area left to the vertical dashed line ($d_{5\%}$=0.0089) shows the 5% smallest *d* values. 86.6% of 4,564 human-mouse orthologous genes have *D* smaller than $d_{5\%}$, suggesting that the detectable expression-profile divergence of 86.6% of genes is lower than the neutral expectation at the 5% significance level. In this figure, we computed the values of d_H and d_M by averaging the Euclidean distances of all possible combinations of probe sets of the same gene, instead of using two randomly picked probe sets as in Figure 1.4.

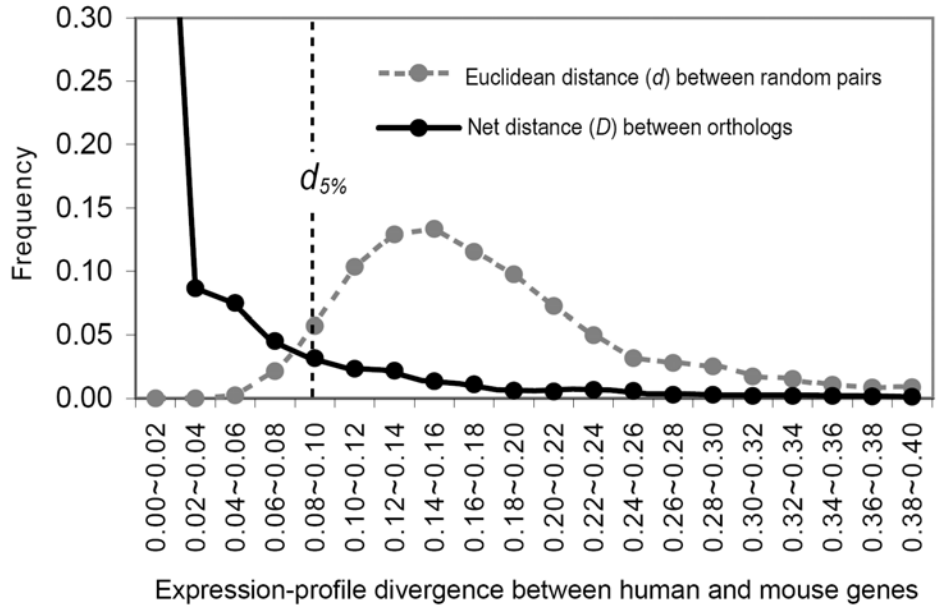


Figure A.3 Correlation between two measures of expression divergence between human-mouse orthologous genes. 10,607 pairs of orthologs are used.

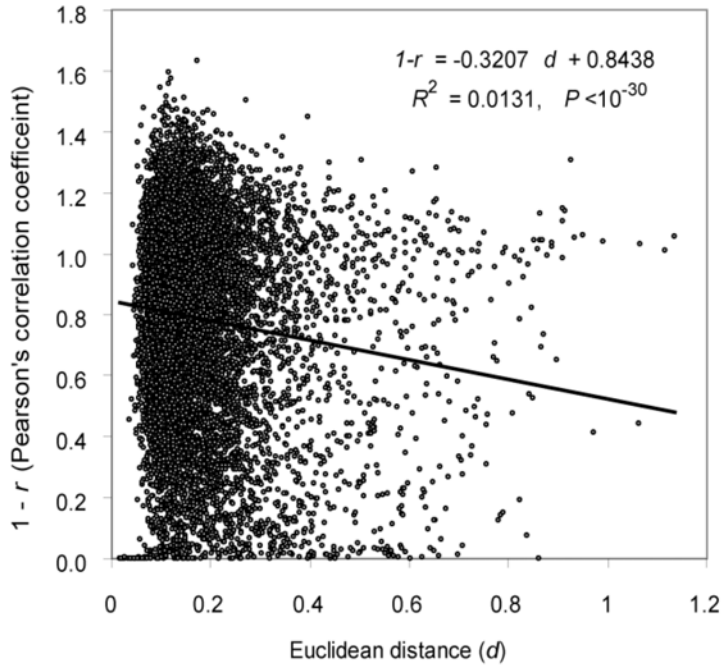


Figure A.4 Tissue-specificity (τ_H) and the coefficient of variation in expression level across tissues (*CV*) are highly correlated (Spearman's rank correlation coefficient = 0.693, $P<10^{-300}$; Pearson's correlation coefficient = 0.690, $P<10^{-300}$). The data are from 10,607 human genes.

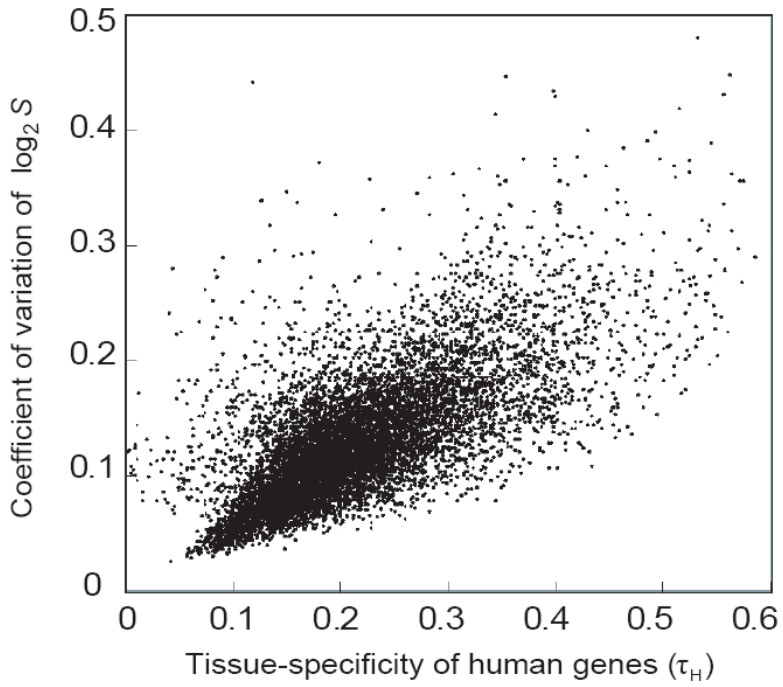


Figure A.5 Highly expressed genes have higher expression-profile similarity between human-mouse orthologs than lowly expressed genes (MAS5 dataset). The expression level is measured by either the mean expression level or the maximum expression level across the 26 common tissues between humans and mice. The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The numbers of genes in each bin are: **(a)** 0-200: 2788, 200-400: 2809, 400-800: 2968, 800-1600: 1441, >1600: 601; **(b)** 0-200: 4441, 200-400: 3166, 400-800: 1977, 800-1600: 730, >1600: 293; **(c)** 0-400: 1762, 400-800: 2599, 800-1600: 2895, 1600-3200: 1777, 3200-6400: 903, >6400: 671; **(d)** 0-400: 3339, 400-800: 2816, 800-1600: 2271, 1600-3200: 1190, 3200-6400: 581, >6400: 410.

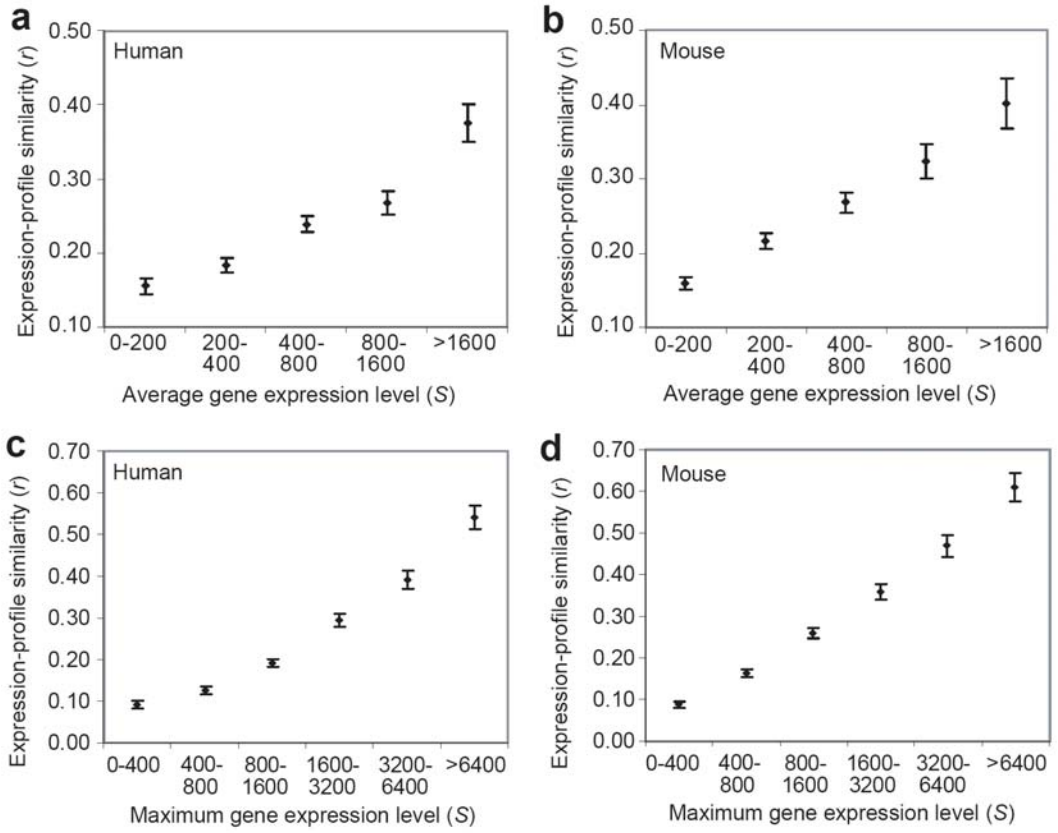


Figure A.6 Highly expressed genes have higher expression-profile similarity between human-mouse orthologs than lowly expressed genes (GC-RMA dataset). The expression level is measured by either the mean expression level or the maximum expression level across all tissues (i.e., 73 human normal tissues or 60 mouse tissue). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The numbers of genes in each bin are: **(a)** 0-200: 4756, 200-400: 1755, 400-800: 1556, 800-1600: 1247, 1600-3200: 775, >3200: 497; **(b)** 0-200: 3823, 200-400: 1512, 400-800: 1581, 800-1600: 1547, 1600-3200: 1118, >3200: 1026; **(c)** 0-400: 2858, 400-800: 1144, 800-1600: 1324, 1600-3200: 1434, 3200-6400: 1432, >6400: 2415; **(d)** 0-400: 2314, 400-800: 883, 800-1600: 1089, 1600-3200: 1365, 3200-6400: 1574, >6400: 3382.

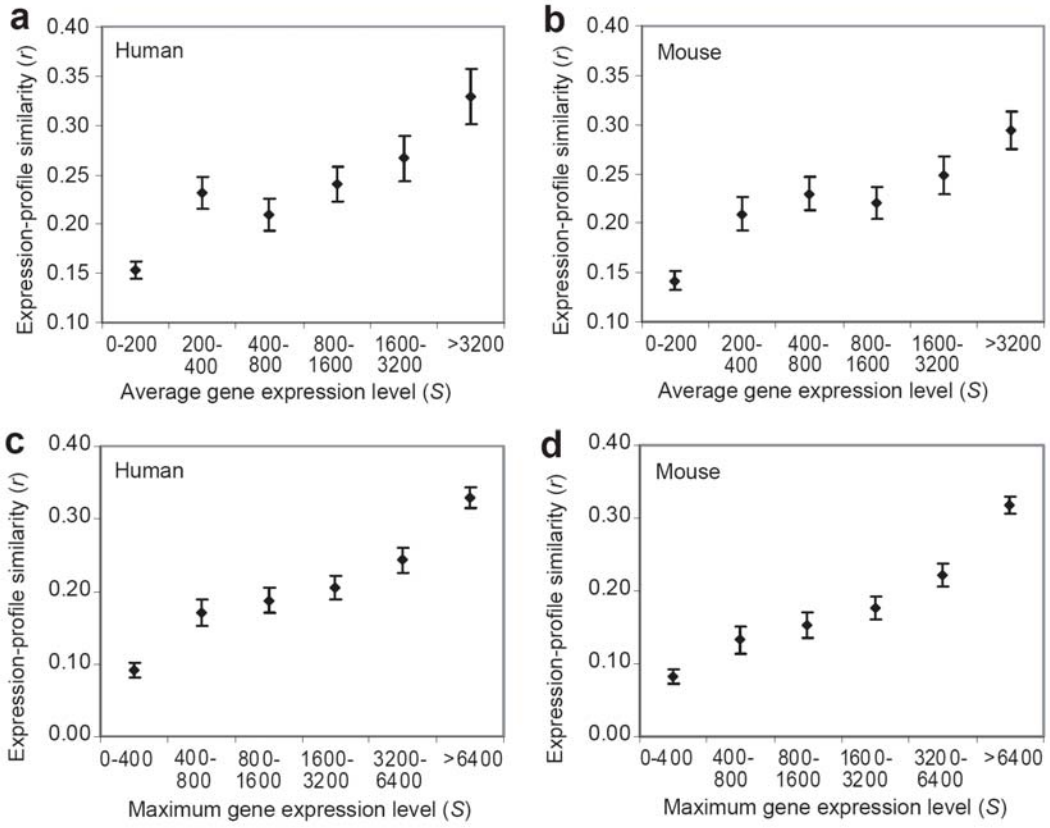


Figure A.7 Greater expression-profile similarities between human-mouse orthologs for genes of high tissue-specificity than genes of low tissue-specificity (GC-RMA dataset). Tissue-specificity is measured using all available tissues (i.e., 73 human normal tissues or 60 mouse tissue). The error bar shows 95% confidence interval of the mean, estimated by 10,000 bootstrap replications for each bin. The numbers of genes in each bin are: (a) 0.00-0.05: 1275, 0.05-0.10: 921, 0.10-0.15: 1203, 0.15-0.20: 1539, 0.20-0.25: 1571, 0.25-0.30: 1403, 0.30-0.35: 990, 0.35-0.40: 724, >0.40: 981; (b) 0.00-0.05: 1155, 0.05-0.10: 940, 0.10-0.15: 1425, 0.15-0.20: 1497, 0.20-0.25: 1312, 0.25-0.30: 1140, 0.30-0.35: 910, 0.35-0.40: 787, >0.40: 1441.

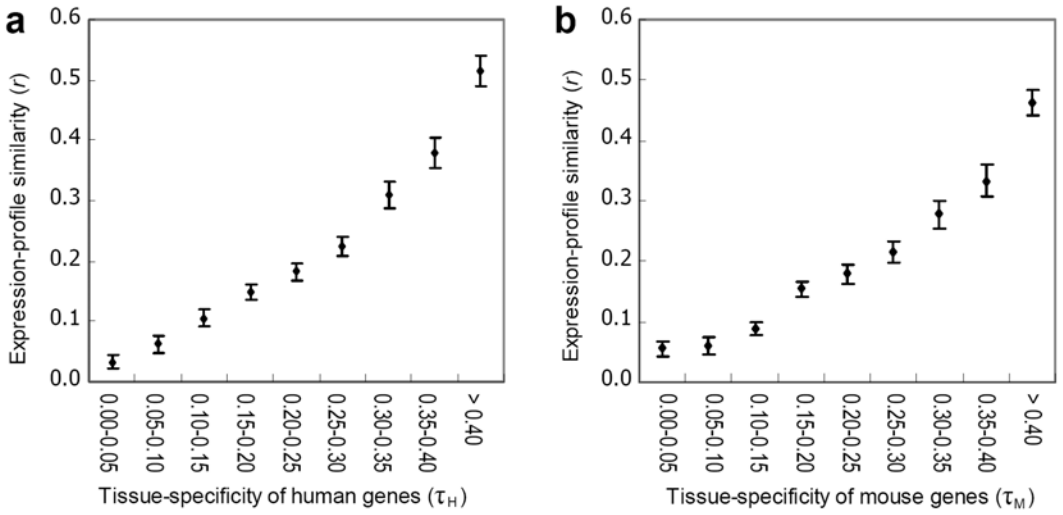


Figure A.8 Expression-profile similarity vs. genomic distance in a randomly permutated genome. Linear regression of average expression-profile similarity of linked genes, measured by *ln*[(1+*R*)/(1-*R*)], versus their $\log_{10}$-transformed genomic distance in nucleotides (log*D*), in a randomly permutated genome. In comparison with the real human genome (Fig 2), there is no correlation between average *ln*[(1+*R*)/(1-*R*)] and log*D***.** The bin size ranges from 20 kilobases (the 1st bin) to ~ 715 kilobases (the last bin) (see Materials and Methods for details on bin sizes).

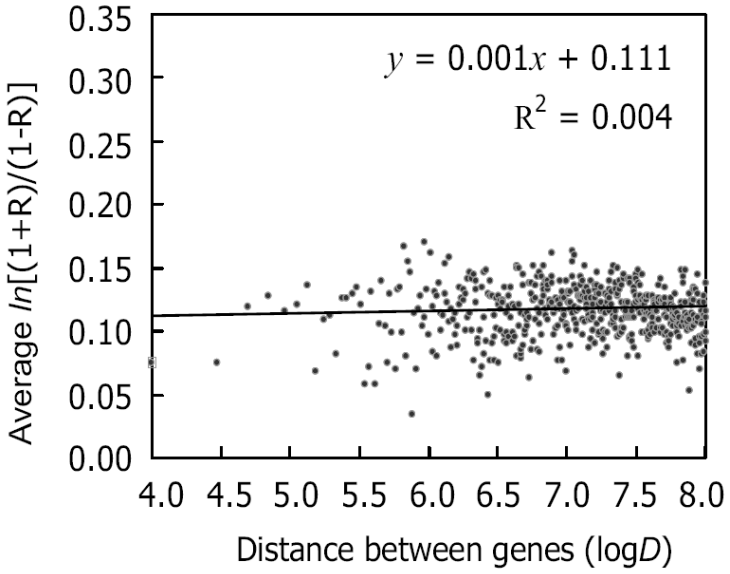


Figure A.9 Expression-profile similarity vs. genomic distance measured by the number of intervening genes. Linear regression of average expression-profile similarity of linked genes, measured by $ln[(1+R)/(1-R)]$, versus their log-transformed genomic distance measured by the number of intervening genes ($\log_2(N)$), where N is set to be the median of each X-axis bin. $ln[(1+R)/(1-R)]$ is strongly negatively correlated with $\log_2 N$**.** Same as in Fig 3, the bin size gradually increases when N becomes larger. The figure is further divided into three areas by gray shading. These three areas are $N<10$, $10<N\leq100$, and $100<N\leq500$, respectively.

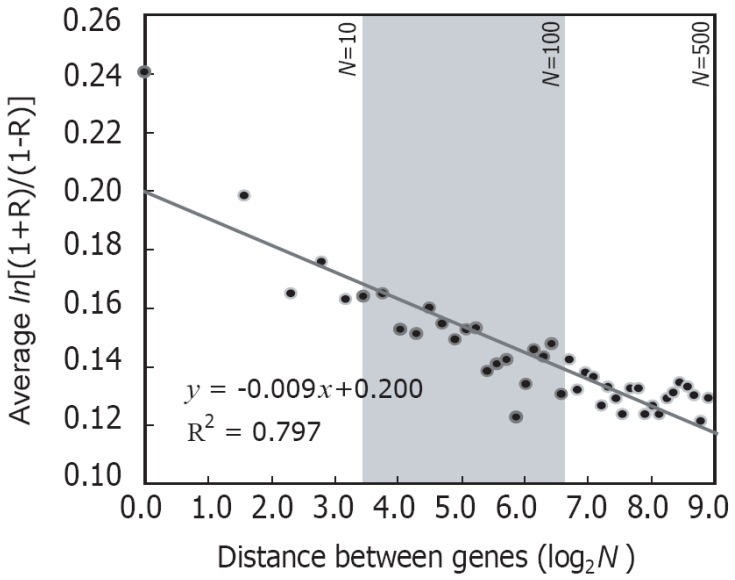


Figure A.10 Linked mouse genes with non-conserved linkage have higher expression-profile similarity than those with conserved linkage. (A) Phylogeny of mouse, human, and dog. Only mouse linked genes that are ancestrally linked, determined by the linkage in dog, are included in the analysis. Black and gray bars represent two genes. Average expression-profile similarity (± standard error), measured by $ln[(1+R)/(1-R)]$, for genes with conserved linkage and genes with non-conserved linkage are shown in **(B)**. The *P* value (two-tailed paired *t*-test) for the hypothesis of no difference in mean expression-profile similarity between genes with conserved linkage and those with non-conserved linkage is 2.1×10^{-2} for **(B)**.

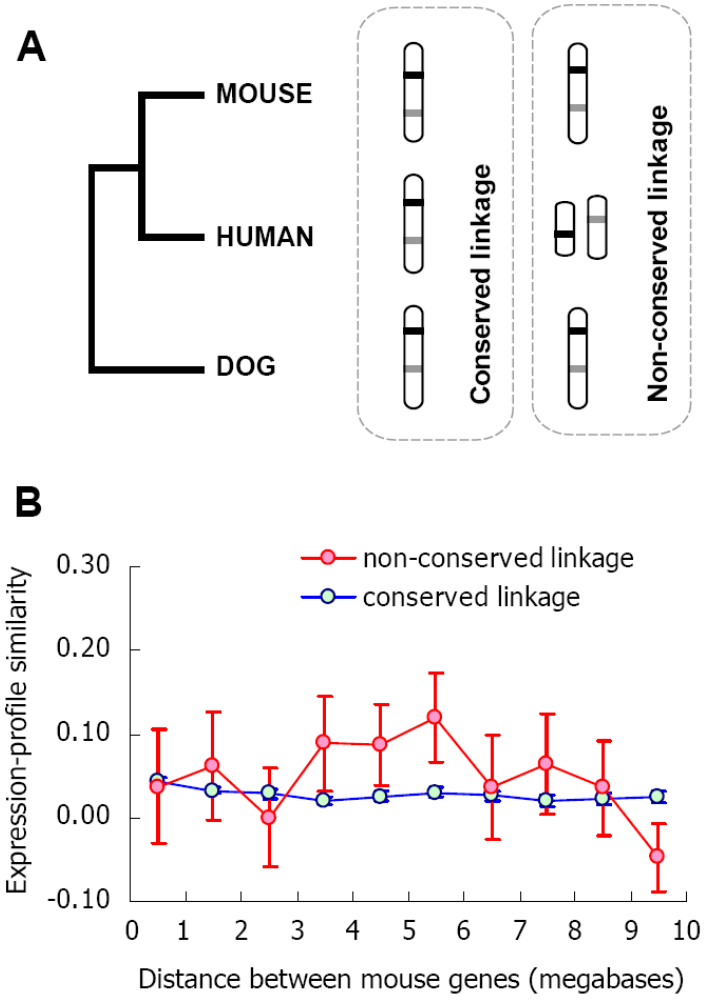


Figure A.11 Higher recombination rates between highly co-expressed genes than poorly co-expressed genes are observed in the human genome after controlling for the chromosomal distance between linked genes by grouping gene pairs with similar *D*. The bin sizes for **(A)** and **(C)** are 0.1Mb and 1Mb, respectively. Highly co-expressed genes are defined by gene pairs with the 10% highest values of $ln[(1+R)/(1-R)]$, whereas poorly co-expressed genes are those with 10% lowest $ln[(1+R)/(1-R)]$, when $ln[(1+R)/(1-R)]$ of every pairs of genes of are computed for each group. Average recombination rates (± standard error) for highly co-expressed genes (black solid circle with black dashed line) and poorly co-expressed genes (empty circle with gray solid line) are shown in **(A)** and **(C)**. **(B)** and **(D)** show the difference in average recombination rates between highly and poorly co-expressed genes (highly minus poorly co-expressed) for **(A)** and **(C)**, respectively. *P* values (two-tailed *t*-test) for the hypothesis that the difference equals to zero are 3.76×10^{-2} and 2.94×10^{-4} for **(B)** and **(D)**, respectively. The analysis is based on 4,857 duplicate-free human autosomal genes. The data of recombination rates across the human genome is based on the map produced by the deCODE project (Kong et al. 2002). The recombination rate (cM/Mb) between two human linked genes was computed by averaging the recombination rates between their positions of transcription starting sites (the genomic regions without recombination data were omitted).

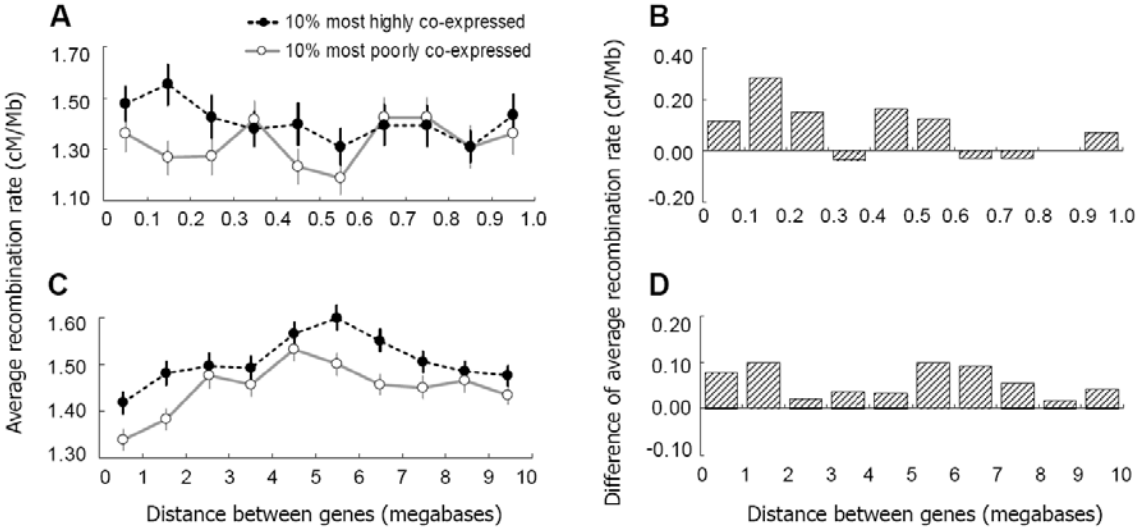


Figure A.12 No correlation between human chromosomal size and average expression-profile similarity of linked genes, measured by *ln*[(1+*R*)/(1-*R*)]. Each dot represents a chromosome. Error bar indicates the standard error. For each panel, the range of distance (size = 2 megabases) between a gene pair (*D*) is shown on the top left corner, while the Spearman's correlation coefficient ρ and associated *P* value are shown on the top right corner.

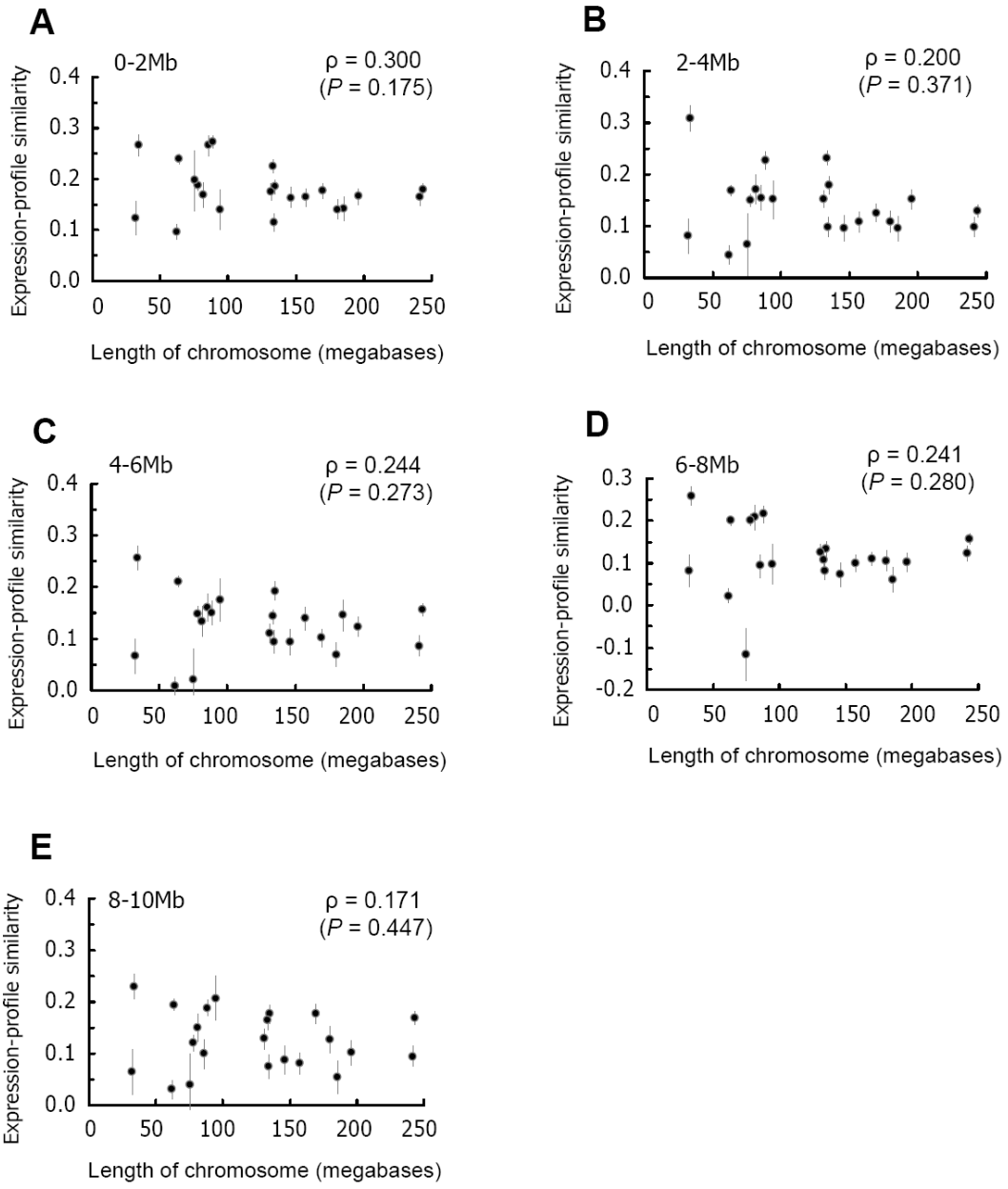


Figure A.13 No correlation between human chromosomal size and average expression-profile similarity of linked genes, measured by *ln*[(1+*R*)/(1-*R*)]. Each dot represents a chromosome. Error bar indicates the standard error. For each panel, the range of distance (size = 5 megabases) between a gene pair (*D*) is shown on the top left corner, while the Spearman's correlation coefficient ρ and associated *P* value are shown on the top right corner.

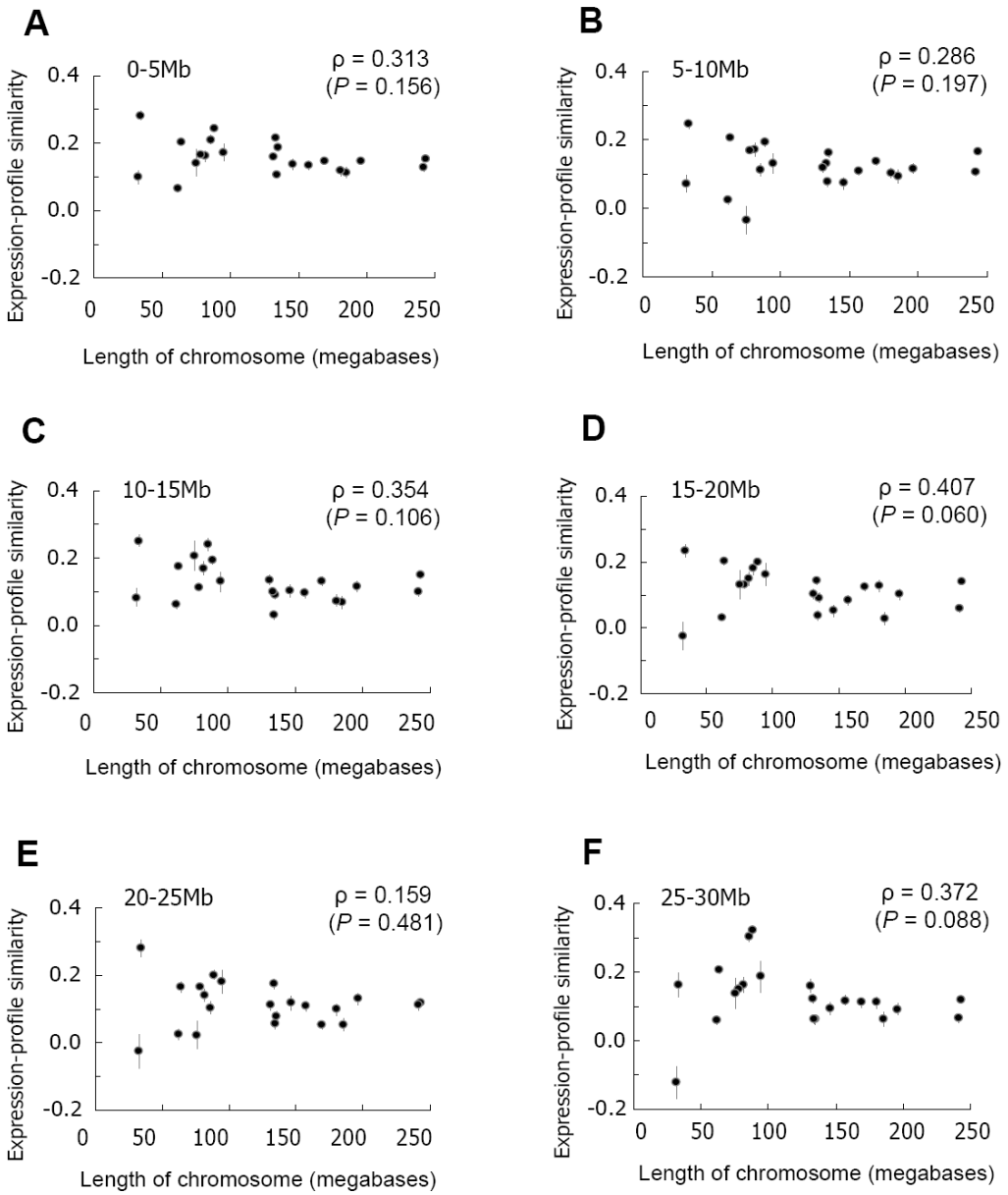


Figure A.14 Expression-profile similarity vs. genomic distance in a single chromosome. Linear regression of average expression-profile similarity of linked genes, measured by $ln[(1+R)/(1-R)]$, versus their $\log_{10}$-transformed genomic distance in nucleotides ($\log D$) within a single chromosome (human chromosome 1 as an example here), where D is set to be the median of each X-axis bin. The result shows that average $ln[(1+R)/(1-R)]$ is strongly negatively correlated with $\log D$. The analysis is based on the gene set free from tandem duplicates.

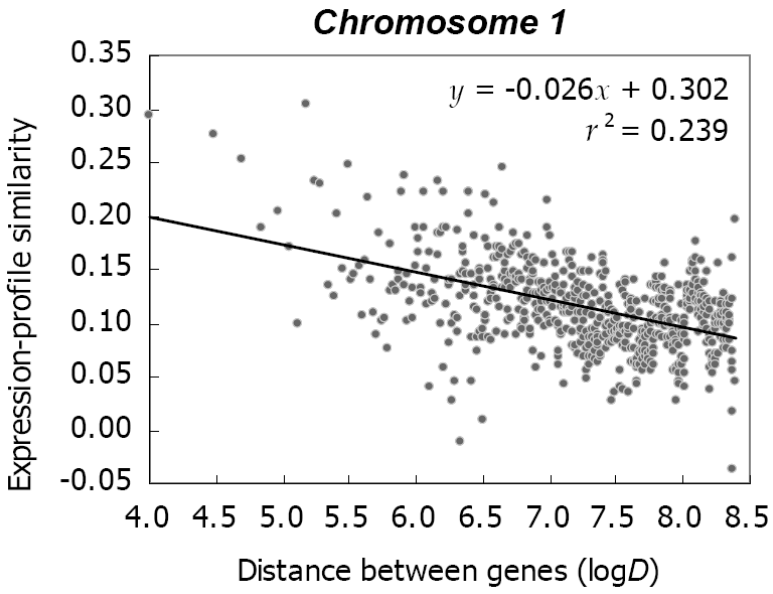


Figure A.15 Number of chromosomal breakage events and the size of the genomic regions covered by co-expressed clusters in the human genome. Open circles represent the numbers from 1,000 randomized human genomes, while the solid triangle indicates the observed numbers from the real human genome. This figure is generated by using the same approach described in Singer et al. (Singer et al. 2005), corresponding to Figure 4C in Singer et al. (Singer et al. 2005). Different from Singer et al. (Singer et al. 2005), we use the updated microarray data (Su et al. 2004), use $ln[(1+R)/(1-R)]$ and 73 tissues to measure the expression-profile similarity, and most importantly, count the chromosomal breakages within co-expressed human clusters that occurred in the mouse lineage after the human-mouse separation. That is, for two consecutive human genes in an identified cluster with known mouse and dog orthologs, if the mouse orthologs are on different chromosomes and the dog orthologs are on the same chromosome (i.e., ancestral linkage), a break event is inferred. The line is the linear regression of the dots. The observed number of breakage events within the real human genomic regions covered by co-expressed gene clusters (triangle) is greater than expected (regression line), suggesting higher rates of linkage breakage within co-expressed gene clusters than in other regions, contrary to Singer et al.'s finding (Singer et al. 2005).

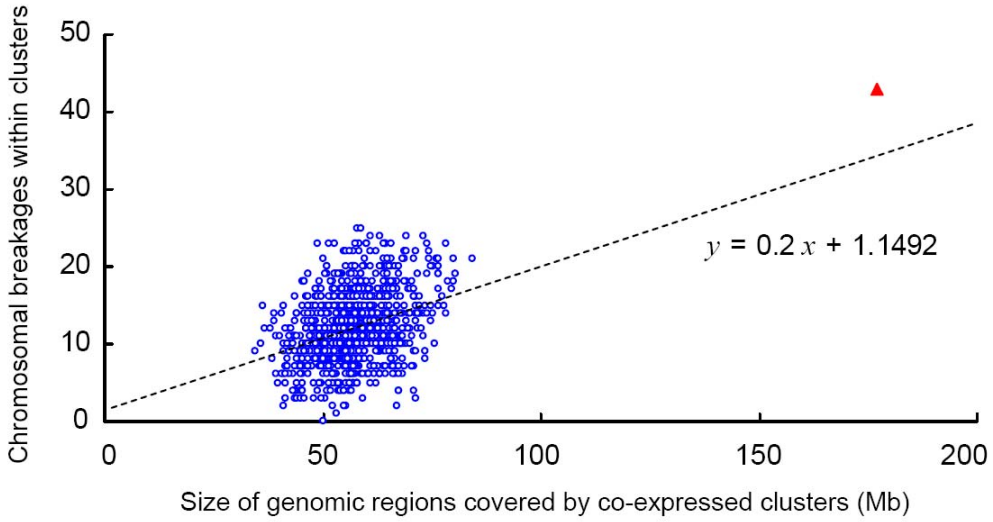


Figure A.16 Expression-profile divergence, measured by 1-*ICE*, between human and mouse orthologous genes (a: ExonArray data, b: GeneAtlas v2 data). *ICE* (index of co-expression) between two genes is defined as the number of tissues in which both genes are expressed divided by the geometric mean of the number of tissues where each gene is expressed (see Methods). Values of upper quartile, median, and lower quartile are indicated in each box. The bars indicate semi-quartile ranges. H and M indicate human and mouse, respectively, and the subscripts e, n, and a indicate essential, nonessential, and any genes, respectively. The *P*-vales are determined by two-tail Mann-Whitney *U* tests.

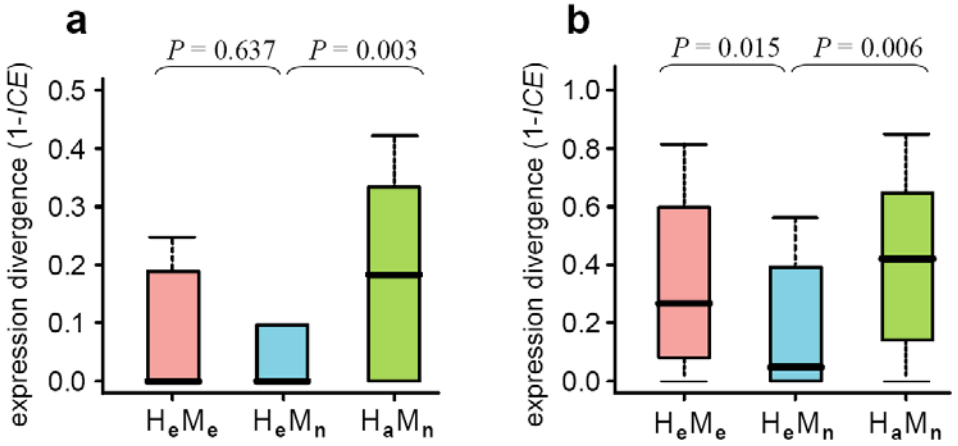


Figure A.17 The quartile-plots of sequence divergence (a: d_N, b: d_S) and expression-profile divergence (c: ExonArray data, d: GeneAtlas v2 data) between human and mouse orthologous genes, after the removal of vacuole proteins. Values of upper quartile, median, and lower quartile are indicated in each box. The bars indicate semi-quartile ranges. H and M indicate human and mouse, respectively, and the subscripts 1, 0, and a indicate essential, nonessential, and any genes, respectively. The *P*-vales are determined by Mann-Whitney *U* tests.

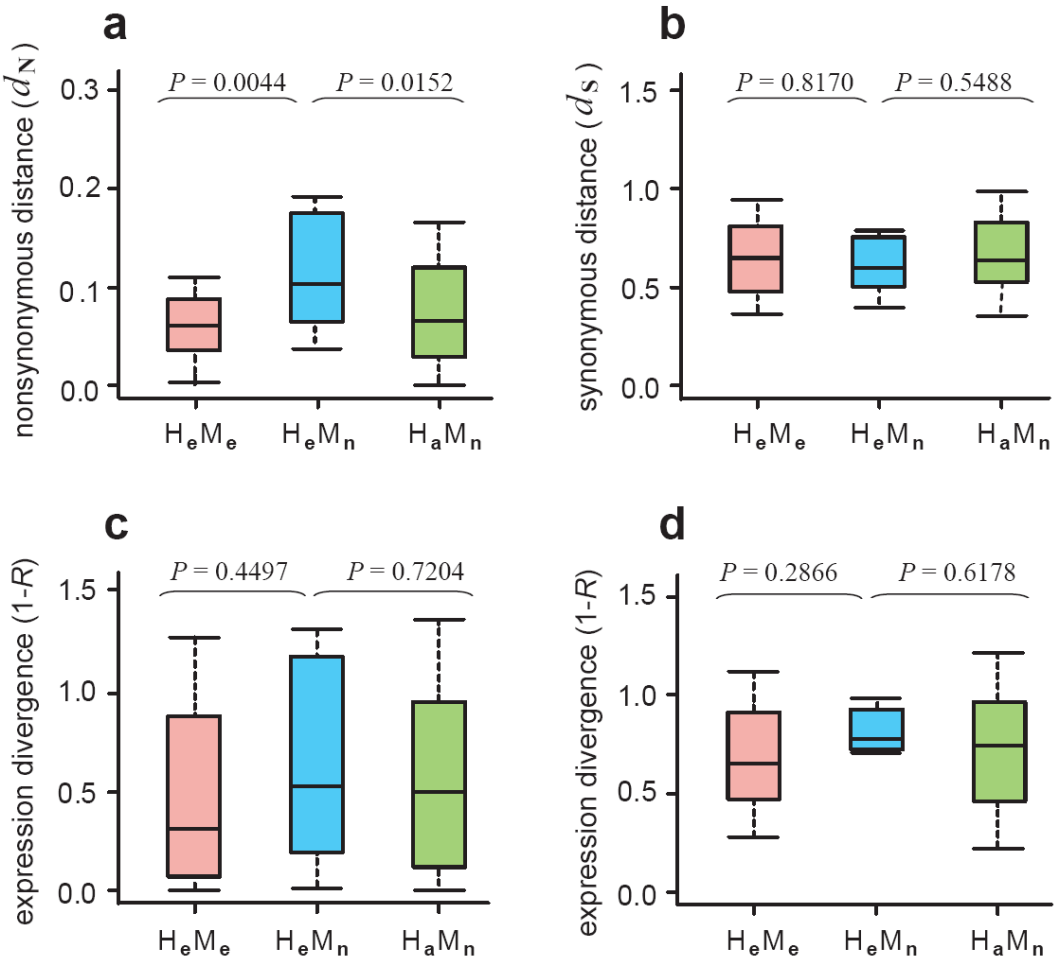


Table A.1 Spearman's rank correlation coefficient (ρ) between gene compactness and d_N, d_S, or d_N/d_S. Gene compactness was calculated based on the shortest isoform when a gene is alternatively spliced. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,575 mouse genes and their rat orthologs, same as in Table 3.1.

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Gene Compactness						
UTR length	-0.196	<2E-16	-0.051	1.04E-02	-0.191	<2E-16
5'-UTR length	-0.126	1.28E-10	-0.055	5.00E-03	-0.116	3.69E-09
3'-UTR length	-0.106	6.21E-08	-0.010	6.26E-01	-0.108	4.37E-08
Intron length (avg)	-0.118	1.77E-09	0.031	1.17E-01	-0.140	8.52E-13

Table A.2 Spearman's rank correlation coefficient (ρ) between gene compactness and d_N, d_S, or d_N/d_S. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 1,955 mouse genes that are not alternatively spliced (a subset of genes in Table 3.1).

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Gene Compactness						
UTR length	-0.234	<2E-16	-0.048	3.30E-02	-0.233	<2E-16
5'-UTR length	-0.158	2.53E-12	-0.045	4.69E-02	-0.154	8.04E-12
3'-UTR length	-0.127	1.87E-08	-0.022	3.33E-01	-0.124	3.87E-08
Intron length (avg)	-0.148	4.28E-11	0.018	4.24E-01	-0.165	2.45E-13

Table A.3 Spearman's rank correlation coefficient (ρ) between gene compactness and d_N, d_S, or d_N/d_S. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 2,354 mouse genes that are not overlapped or nested with any other gene in the mouse genome (a subset of genes in Table 3.1).

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Gene Compactness						
UTR length	-0.210	<2E-16	-0.044	3.18E-02	-0.209	<2E-16
5'-UTR length	-0.140	8.34E-12	-0.056	6.23E-03	-0.129	2.94E-10
3'-UTR length	-0.102	6.40E-07	-0.008	6.96E-01	-0.106	2.88E-07
Intron length (avg)	-0.172	<2E-16	-0.024	2.42E-02	-0.181	<2E-16

Table A.4 Spearman's rank correlation coefficient (ρ) between gene compactness and d_N, d_S, or d_N/d_S. *P*-values show the probabilities of the observations under the hypothesis of no correlation. The analysis is based on 17,465 mouse-rat orthologous genes.

	d_N		d_S		d_N/d_S	
	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Gene Compactness						
UTR length	-0.243	<2E-16	-0.115	<2E-16	-0.224	<2E-16
5'-UTR length	-0.182	<2E-16	-0.104	<2E-16	-0.162	<2E-16
3'-UTR length	-0.185	<2E-16	-0.089	<2E-16	-0.169	<2E-16
Intron length (avg)	-0.136	<2E-16	-0.040	1.25E-07	-0.134	<2E-16

Table A.5 Correlation between chromosomal distance (log*D*) and average expression-profile similarity, measured by *ln*[(1+*R*)/(1-*R*)], between mouse linked genes.

Genomic distance (*D*)	Pearson's *r*	Chance probability
<1Mb	-0.5597	< 0.001
1-5Mb	0.0717	0.753
5-25Mb	-0.2433	0.002
25-50Mb	-0.1322	0.122
50-100Mb	0.0129	0.838

Table A.6 Comparison between the mouse genes from the H_eM_n group and those from the H_eM_e group.

Minimal protein sequence identity in defining paralogs	60%	70%	80%	Not required
Proportion of mouse genes that have paralogs (H_eM_n, H_eM_e; *P*-value from Fisher's exact test [a])	29.6% (8/27), 29.0% (27/93); *P*=1.000	11.1% (3/27), 16.1% (15/93); *P*=0.568	3.7% (1/27), 5.4% (5/93); *P*=1.000	66.7% (18/27), 59.1% (55/93); *P*=0.512
Average number of paralogs [b] (H_eM_n, H_eM_e; *P*-value from the *U* test [a])	2.13, 1.48; *P*=0.787	2.66, 1.26; *P*=0.097	1.00, 1.00; *P*=1.000	4.33, 3.78; *P*=0.415
Average protein sequence identity to the closest paralog [b] (H_eM_n, H_eM_e; *P*-value from *U* test [a])	66.8%, 72.9%; *P*=0.098	73.7%, 78.2%; *P*=0.190	80.0%, 85.4%; *P*=0.333	56.2%, 58.3%; *P*=0.568

[a] The null hypothesis is equal values between the H_eM_n and H_eM_e groups. Two-tail tests are conducted.

[b] Only mouse genes that have paralog(s) are counted.

Table A.7 Basal metabolic rates (*BMR*) and reproductive ages (*T*) of primates and several other mammals. *BMR* × *T* refers to the relative amount of metabolic waste generated per gram of body mass until reproduction.

Species name (common name)	Mean *BMR* [a] (cal per gram body mass per day)	Reproductive age *T* [b] (year)	*BMR* × *T*
Homo sapiens (human)	23.6	20.0[c]	472.0
Pan troglodytes (chimpanzee)	27.9	13.5[d]	376.7
Gorilla gorilla (gorilla)	19.7	15.0[d]	295.5
Pongo pygmaeus (orangutan)	35.1	12.0[e]	421.2
Papio anubis (olive baboon)	43.2	8.5[f]	267.2
Macaca mulatta (Rhesus monkey)	37.0	4.5[d]	166.5
Chlorocebus aethiops (green monkey)	43.4	5.0[d]	216.5
Saguinus mystax (mustached tamarin)	88.4	1.5[d]	132.6
Otolemur crassicaudatus (greater galago)	68.4	2.0[d]	136.8
Eulemur fulvus (brown lemur)	57.6	1.5[d]	86.4
Tupaia glis (common tree shrew)	100.0	0.25[d]	25.0
Peromyscus maniculatus (deer mouse)	151.0	0.13[d]	19.6
Mus musculus (house mouse)	189.0	0.14[d]	26.5

[a] BMR of young adult males under resting/fasting condition (Tolmasoff et al. 1980. *Proc. Natl. Acad. Sci. USA* 77:2777-2781)
[b] Age at onset of male reproduction.
[c] Fenner. 2005. *Am. J. Phys. Anthropol.* 128:415-423.
[d] Animal Diversity Web (http://animaldiversity.ummz.umich.edu/).
[e] Macdonald. 2001. *The Encyclopedia of Mammals*. Andromeda Oxford Ltd., Abingdon, UK.
[f] Strum and Western. 1982. *Am. J. Primatol.* 3:61–76.