Coexpression of Linked Genes in Mammalian Genomes Is Generally Disadvantageous

Ben-Yang Liao and Jianzhi Zhang

Department of Ecology and Evolutionary Biology, University of Michigan

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 coexpression patterns. The adaptive model assumes that coexpression is advantageous and is established by relocation of initially unlinked but coexpressed genes, whereas the neutral model asserts that coexpression 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 coexpression 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 coexpression in the human genome. However, the linkage is more likely to be broken during evolution between genes of high coexpression than those of low coexpression and the breakage of linkage reduces gene coexpression. These results support our hypothesis that coexpression 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.

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 et al. 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 et al. 2003; Versteeg et al. 2003), broadly expressed genes (Lercher et al. 2002), and coexpressed genes (Fukuoka et al. 2004; Singer et al. 2005; Semon and Duret 2006) have been observed in the human genome. The general phenomenon of coexpression of linked genes has also been reported in other model eukaryotes such as the yeast *Saccharomyces cerevisiae* (Coghlan and Wolfe 2000; Kruglyak and Tang 2000; Huynen et al. 2001; Fukuoka et al. 2004; Lercher and Hurst 2006), nematode *Caenorhabditis elegans* (Lercher et al. 2003; Fukuoka et al. 2004), and fruit fly *Drosophila melanogaster* (Boutanaev et al. 2002; Spellman and Rubin 2002; Bailey et al. 2004; Fukuoka et al. 2004; Kalmykova et al. 2005).

However, it is unclear as to how and why linked genes become coexpressed. 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 coexpression of linked genes may be important to gene function (Hurst et al. 2002; Singer et al. 2005). This view, referred to as the adaptive model, assumes that it is beneficial for genes that require coexpression 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 coexpressed gene cluster is established, the linkage of the coexpressed genes should be evolutionarily maintained by purifying selection (Hurst et al. 2002; Singer et al. 2005).

Key words: gene order, linkage, gene expression, coexpression, evolution, mammals.

E-mail: jianzhi@umich.edu.

Mol. Biol. Evol. 25(8):1555–1565. 2008
doi:10.1093/molbev/msn101
Advance Access publication April 24, 2008

Observations of functional similarity of coexpressed 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 coexpressed neighboring genes (Spellman and Rubin 2002). In humans, clusters of coexpressed linked genes that belong to the same functional category, as defined by GO, are rare (Fukuoka et al. 2004). Furthermore, although the evolutionary conservation of linkage between coexpressed genes in several yeasts supports the adaptive model (Hurst et al. 2002), considering the recent discovery of long-range coregulation (~100 kb, 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 et al. 2004; Sproul et al. 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 coexpressed linked genes suggested that coexpressed 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 et al. 2005). The neutral model for the formation of coexpressed 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 et al. 2004), these studies were later shown to have either technical problems or alternative interpretations (Liao and Zhang 2006a). 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 et al. 2005; Khaitovich et al. 2005; Rifkin et al. 2005; Liao and Zhang 2006a; Whitehead and Crawford 2006; Xing et al. 2007). Because coexpression 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 coexpressed gene clusters. Here, we propose that coexpression 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 data set (Su et al. 2004). We examined coexpression 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 coexpression than those with low coexpression, consistent with the predictions of our hypothesis. Based on these findings, we propose a model of the origin and evolutionary dynamics of coexpression of linked genes.

Materials and Methods

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 unique best reciprocal hit (UBRH, meaning that they were UBRHs 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.

Analysis of the Microarray Data

We obtained the expression information of human genes and mouse genes from the Gene Atlas V2 data set (http://symatlas.gnf.org/SymAtlas/) (Su et al. 2004). This data set 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 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 et al. 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 2006a), we arbitrarily chose the probe set with highest expression level (Jordan et al. 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.

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 2 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 coexpression pattern artificially generated by cross-hybridization.

We followed the conventional approach (Lercher et al. 2002, 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 data set of a genome (for genes having more than one isoforms, 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 et al. 2002). We then generated a duplicate-free data set 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 $\sim$25% (from 4,857 to 3,681). To maintain the statistical power and keep our data set representative of the whole genome, we generated a tandem duplicate–free data set, which contains 7,577 human genes that have expression data and have orthologs in the other 3 mammalian genomes. This data set is larger than the above duplicate-free data set because we now allow duplicate genes that are located on different chromosomes.

Expression Profile Similarity between Linked Genes

Following Gu et al. (2002), we measured the level of coexpression between 2 linked genes (say A and B)

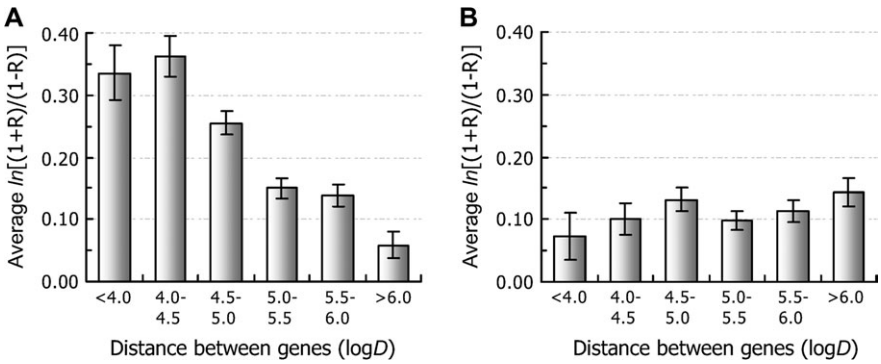

FIG. 1.—Expression profile similarities between adjacent human genes, measured by $ln[(1 + R)/(1 - R)]$, are negatively correlated with log$D$, their $\log_{10}$-transformed genomic distance in nucleotides, in (*A*) the real human genome, but not in (*B*) the permutated human genome. Average $ln[(1 + R)/(1 - R)]$ (±standard error) is shown for each group of adjacent genes categorized by log$D$. The number of gene pairs per category is 213, 488, 966, 1,343, 1,111, and 714, respectively, for the 6 categories.

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 coexpression. 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 2 genes, as annotated by Ensembl.

In figure 3 and supplementary figure S3 (Supplementary Material online), the size of each bin was fixed to a certain value. In figure 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 nth bin is $10^6 \times 2^{(99+n)/100}$nt. That is, the nth 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 1×10^6 to 2×10^6. Use of other bin sizes did not change our results qualitatively.

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 2 genes are located in the same chromosome. Although it is possible that 2 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 nonconserved, meaning that the linkage is lost in one or both rodents.

Results

Coexpression of Distantly Linked Human Genes

It is important to first examine whether the phenomenon of coexpression 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 et al. 2004). Some studies have attempted to address this question by examining the on/off expressional status of linked genes (Lercher et al. 2002; Semon and Duret 2006), whereas other studies examined the correlation of across-tissue expression profiles of adjacent genes (Hurst et al. 2002; Singer et al. 2005). Adjacent genes are linked genes without any other genes in between. Because chromosomal rearrangements between sex chromosomes and autosomes are rare and sex-linked genes have special functions and expression profiles (Lahn et al. 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 coexpression ($ln[(1 + R)/(1 - R)]$, 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 (fig. 1*A*). In comparison, the human genome with permutated expression profiles, which is generated by randomly assigning gene names to the real expression profiles (of the 4,835 duplicate-free genes), shows no obvious pattern between log$D$ and $ln[(1 + R)/(1 - R)]$ (fig. 1*B*). Our observations are consistent with previous studies in the yeast (Hurst et al. 2002).

The power to decipher the effect of linkage on gene coexpression 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 data set). 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

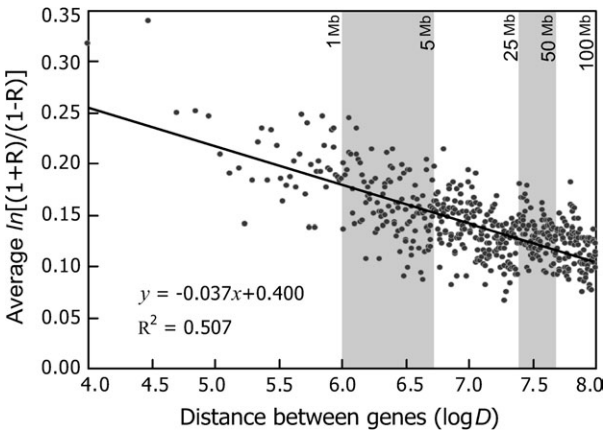

FIG. 2.—Linear regression of average expression profile similarity of linked genes, measured by $ln[(1 + R)/(1 - R)]$, versus their $\log_{10}$-transformed genomic distance in nucleotides (log$D$), where D is set to be the median of each x axis bin. In the real human genome, average $ln[(1 + R)/(1 - R)]$ is strongly negatively correlated with log$D$. The bin size ranges from 20 kb (the 1st bin) to ~715 kb (the last bin) (see Materials and Methods for details on bin sizes). The figure is further divided into 5 areas by gray shading. These 5 areas are <1, 1–5, 5–25, 25–50, and 50–100 Mb. The correlations between $ln[(1 + R)/(1 - R)]$ and log$D$ of these 5 areas are shown in table 1.

Table 1
Correlation between Chromosomal Distance (log$D$) and Average Expression Profile Similarity, Measured by $ln[(1 + R)/(1 - R)]$, between Human-Linked Gene Pairs

Genomic Distance (D)	Pearson's r	Chance Probability
<1 Mb	−0.5808	<0.001
1–5 Mb	−0.4523	<0.001
5–25 Mb	−0.4416	<0.001
25–50 Mb	−0.2693	0.069
50–100 Mb	−0.2391	0.119

NOTE.—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 figure 2.

with the genomic distances ranging up to 100 Mb. 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 $\rho = -0.6227$, $P < 10^{-56}$; fig. 2). On the contrary, the genome with permutated expression profiles shows no correlation (Pearson's $r = 0.0198$, $P = 0.654$; Spearman's $\rho = -0.0700$, $P = 0.1122$; supplementary fig. S1, Supplementary Material online). Because the correlation observed in the real human genome is computed from the data points with D varying from 10 kb to 100 Mb (fig. 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 5 categories based on the value of D: <1, 1–5, 5–25, 25–50, 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 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 <1, 1–5, and 5–25 Mb (table 1), indicating that the phenomenon of coexpression 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 2 linked genes by the number (N) of intervening genes between them. Consistent with figure 2, the correlation between N in $\log_2$ scale and $ln[(1 + R)/(1 - R)]$ is significantly negative (supplementary fig. S2, Supplementary Material online), indicating that our observation does not depend on how the distance is measured and that linked genes with >100 intervening genes are still significantly coexpressed.

To examine whether the phenomenon of long-range gene coexpression 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 $D < 1$ and 5–25 Mb, the negative correlations do not exist for the groups of 1–5, 25–50, and 50–100 Mb in mouse (supplementary table S1, Supplementary Material online).

Weaker Evolutionary Conservation of Linkage between Genes of Higher Coexpression

If the long-range coexpression 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 coexpression of linked genes is to measure the evolutionary conservation of linkage. If coexpression of linked genes is favored by natural selection, the linkage should be maintained during evolution. If coexpression 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 coexpression and those with low levels of coexpression. If coexpression of linked genes is detrimental, the linkage of highly coexpressed genes should be broken more often during evolution than that of poorly coexpressed 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 figure 3*A* (Springer et al. 2003; Murphy et al. 2004; Kriegs et al. 2006; Nishihara et al. 2006), we infer that 2 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 et al. 2006), the phylogeny we use here has been well established by analyses of both irreversible genomic events (Kriegs et al. 2006; Nishihara et al. 2006) and DNA sequences from many taxa (Springer et al. 2003; Murphy et al.

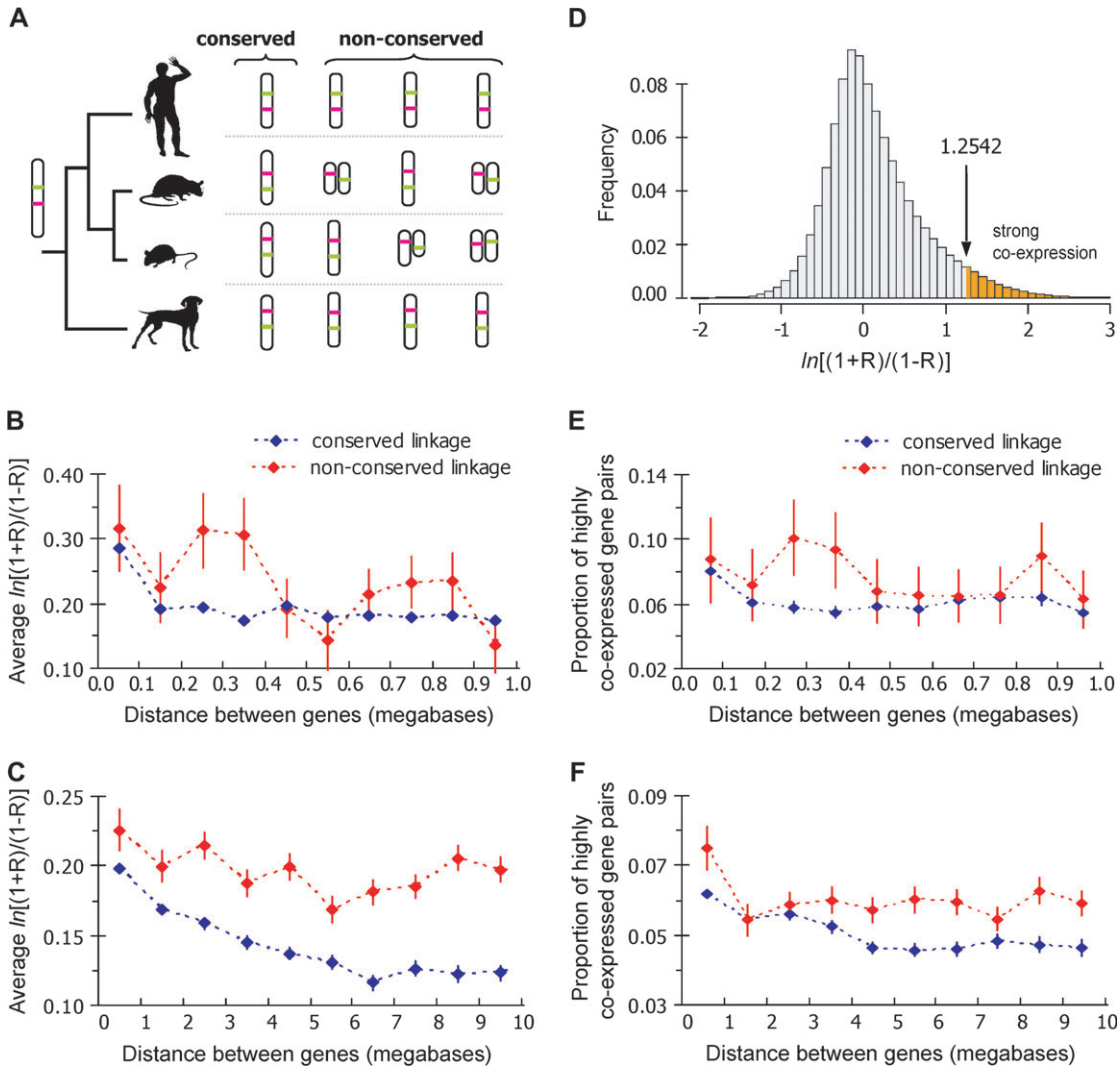

FIG. 3.—Linked human genes with nonconserved linkage have higher expression profile similarity than those with conserved linkage. (*A*) Phylogeny of human, rat, mouse, and dog. Only human-linked genes that are ancestrally linked, determined by the linkage in dog, are included in the analysis. Pink and green bars represent 2 genes. (*B*) Average expression profile similarity (±standard error), measured by $ln[(1 + R)/(1 - R)]$, for genes with conserved linkage and genes with nonconserved linkage, at short physical distances. (*C*) Average expression profile similarity for genes with conserved linkage and genes with nonconserved linkage, at long physical distances. (*D*) Distribution of $ln[(1 + R)/(1 - R)]$ for all linked (duplicate-free) gene pairs. Strongly coexpressed 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. (*E*) Average expression profile similarity (±standard error) for strongly coexpressed genes with conserved linkage and genes with nonconserved linkage, at short physical distances. (*F*) Average expression profile similarity for strongly coexpressed genes with conserved linkage and genes with nonconserved linkage, at long 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 nonconserved 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.

2004) and thus is much more trustable 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 (fig. 3*A*). Because rodent genomes have gone through extensive rearrangements during evolution (Bourque et al. 2004; Mullins LJ and Mullins JJ 2004), current organizations of mouse and rat genomes help divide these ancestrally linked genes into 2 groups: genes with conserved linkage and genes with nonconserved linkage. We then compare the level of coexpression between gene pairs with conserved linkage and those with nonconserved linkage. Because genomic distance *D* influences expression similarity (figs. 1 and 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 nonconservatively linked genes within each group. The results show that, for nearly every *D* range, nonconservatively linked human genes have a higher degree of coexpression than conservatively linked human genes (fig. 3*B* and *C*). This finding is inconsistent with the adaptive model (Hurst et al. 2002; Singer et al. 2005) and the neutral model (Semon and Duret 2006) but is predicted by our hypothesis that coexpression of linked genes is generally detrimental and disfavored by natural selection. We also compare the expression similarity

between conservatively and nonconservatively linked gene pairs when we define nonconservation by a loss of linkage in primates, instead of rodents. The results (supplementary fig. S3, Supplementary Material online) are similar to those in figure 3*B* and *C*, suggesting that the phenomenon of weaker evolutionary conservation of linkage between genes of higher coexpression 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 coexpression (or interference) only acts on weakly to moderately coexpressed linked genes but not on strongly coexpressed linked genes. To define strongly coexpressed 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 3*A–C*) and considered linked genes with $ln[(1 + R)/(1 - R)]$ values falling within the top 5% of the distribution (fig. 3*D*) to be strongly coexpressed. Interestingly, we found that the proportion of strongly coexpressed gene pairs is lower among those with conserved linkage than with nonconserved lineage (fig. 3*E* and *F*), suggesting natural selection against the conservation of linkage of strongly coexpressed gene pairs.

Our transcriptional interference hypothesis predicts that the breakage of linkage between 2 genes would reduce the degree of their coexpression. 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). Because coexpression of linked genes is much weaker in mouse than in human (supplementary table S1, Supplementary Material online), there is a general trend of reduction in expression profile similarity between a gene pair in mouse compared with that in human (fig. 4). However, the reduction is greater for the gene pairs that experienced interchromosomal rearrangements than those that did not (fig. 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) and high recombination tends to disrupt gene linkage (Poyatos and Hurst 2006). Therefore, one expects to observe lower recombination rates between highly coexpressed genes than between poorly coexpressed genes, if coexpression of linked genes is beneficial. However, our analysis of the human genome shows that highly coexpressed linked genes actually have higher recombination rates (cM/Mb) than poorly coexpressed linked genes (supplementary fig. S4, Supplementary Material online). 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 coexpression of linked genes is detrimental.

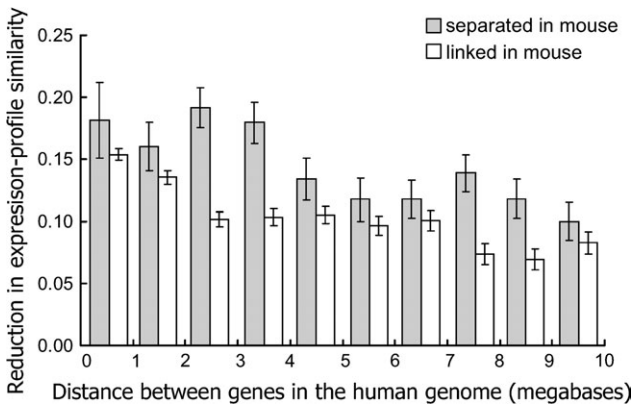

FIG. 4.—Coexpression of linked genes is reduced by interchromosomal rearrangements. Only human-linked genes that are ancestrally linked, determined by the linkage in dog, are included in the analysis. Mouse orthologs of these ancestrally linked genes can be either linked on the same chromosome (white bars) or separated on different chromosomes (black bars). The *y* axis shows the difference in expression profile similarity (±standard error), measured by $ln[(1 + R)/(1 - R)]$, of 2 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 2 groups of genes is 7.83×10^{-4}.

Discussion

There are generally 3 molecular mechanisms that could cause the coexpression of linked genes (Hurst et al. 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 cosuppress the transcriptional activities of several linked genes until reaching boundary elements (Labrador and Corces 2002). This type of coregulation affects regions of up to a few hundred kilobases. At the tertiary level, transcriptional coregulation can happen in 2 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 T and Cremer C 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) coexpression 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 coexpression 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). Because the expression data used here are 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 coexpression of linked genes on different chromosomes. If the level of coexpression is higher in small chromosomes than in large chromosomes for a given D, the results of figure 2 and table 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 coexpression and chromosomal size when control for D (supplementary figs. S5 and S6, Supplementary Material online). Moreover, the negative correlation between the level of gene coexpression and physical distance within a single chromosome is similar to the genome-wide pattern (supplementary fig. S7, Supplementary Material online). It is worth mentioning that one yeast study proposed that the seemingly long-range coexpression 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 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 2, suggesting that our results are not due to special genes in subtelomeric regions. We conclude that the long-range coexpression 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 figures 1 and 2 suggest the existence of these 2 levels of regulations as well.

Contrary to the hypothesis that coexpressed gene clusters correspond to large chromatin domains (Hurst et al. 2002; Roy et al. 2002; Hurst et al. 2004; Sproul et al. 2005), a recent study showed that coexpression of mammalian genes is mainly due to the coregulation of 2 genes by shared promoters (Semon and Duret 2006). Our result favors the hypothesis of gene coregulation by large domains, which is consistent with the discovery in yeast (Lercher and Hurst 2006). Different from our approach Semon and Duret (2006) followed the method used in Lercher et al. (2002) to measure the expression profile similarity of 2 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 (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 coexpression of 2 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 2006b).

Previous investigators have used evolutionary conservation of linkage to study the potential adaptive value of linkage of coexpressed genes, but they did not use outgroups to separate the formation of new linkages from the breakage of old linkages (Hurst et al. 2002; Singer et al. 2005; Semon and Duret 2006; Poyatos and Hurst 2007; Ranz et al. 2007). Hence, if a pair of highly coexpressed 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 2 species. These 2 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 interchromosomal rearrangements during rodent evolution for gene pairs with high coexpression in humans than those with low coexpression (fig. 3). Therefore, coexpression 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. (2005) by counting the interchromosomal breakages within clusters (see fig. 4*C* in Singer et al. [2005]) that occurred in the mouse lineage after the divergence of primates and rodents. The new result (supplementary fig. S8, Supplementary Material online) is opposite of result of Singer et al. and becomes consistent with our findings in figure 3.

Our observations suggest no adaptive value for clustering of coexpressed genes in the human genome in general. Rather, linked genes are coexpressed 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 coexpressed genes may represent lineage-specific transient adaptations (Poyatos and Hurst 2007; Ranz et al. 2007). Although this scenario remains possible, it is extremely hard to test by comparative approaches. Furthermore, this scenario is not contradictory to our finding that coexpression 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 et al. 2003; Hurst et al. 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 coexpression (Alexeyenko et al. 2006). Together with our finding, it is clear that the phenomenon of coexpression 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 coexpression of linked genes is detrimental raises an important question. That is, if such coexpression 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 coexpression of linked genes is generally detrimental, the "mutation" that generates coexpression as a by-product may initially be advantageous. Figure 5 shows an example explaining this model. For simplicity, only 2 genes, A and B, are shown. Initially, A and B are linked but with distinct expression patterns (fig. 5*A*). However, the expression of B is not optimized. When a mutation occurs to establish a transcriptional background for the 2 genes, they become coexpressed. This mutation makes the expression pattern of B closer to its optimal, whereas the coexpression makes the expression pattern of A deviate from its optimal (fig. 5*B*). 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 (fig. 5*C*). It is possible that many genes are involved in a similar evolutionary process as shown in figure 5 because the mechanism creating the transcriptional background has long-range effects. The above verbal model lacks many quantitative details because the molecular mechanism responsible for coregulation of linked genes is poorly known. In the future, when the molecular mechanism of coregulation 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 (fig. 3). Other mechanisms, such as the increase of intergenic distance (Byrnes et al. 2006) and establishment of insulators (Bell et al. 2001), have also been reported. We found the phenomenon of long-range coexpression of linked genes to be much more prominent in human than in mouse (table 1 and supplementary table S1, Supplementary Material online), consistent with the earlier observation that short-range coexpression 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 high background noises compared with the human data, resulting in weaker coexpression 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 coexpression of linked genes in the mouse genome because rearranged mouse orthologs of human-linked genes have a greater reduction in expression

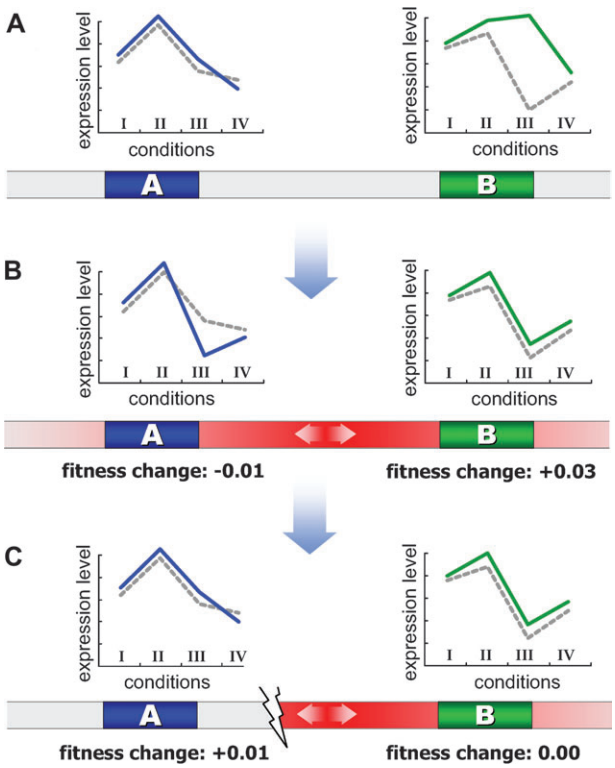

FIG. 5.—The birth and breakdown of coexpression of linked genes. (*A*) The initial expression status of gene A and gene B. Gene A and gene B are linked but not coexpressed. The expression profiles of A and B are shown above the boxes representing the genes. Solid lines represent the current expression 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 coexpression 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 causes coexpression of A and B and drives the expression profile of B closer to but that of A away from their respective optimal expression profiles. 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 coexpression 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 coexpressed. Because the rearrangement increases the overall fitness, this mutation can be fixed.

profile similarity than nonrearranged mouse orthologs (fig. 4). However, because large conserved syntenic blocks (>50 Mb) still exist between human and mouse and the total number of syntenic blocks is no more then 400 (Waterston et al. 2002; Bourque et al. 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.

Conclusions

Our observations presented in the present study are consistent with neither the adaptive nor the neutral model. The results support our hypothesis that coexpression 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 coregulations that are beneficial to the organisms.

Supplementary Material

Supplementary figures S1–S8 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

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.

Literature Cited

Akhunov ED, Akhunova AR, Linkiewicz AM, et al. (31 co-authors). 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. Proc Natl Acad Sci USA. 100:10836–10841.

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

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

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

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

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

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

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

Cannarozzi G, Schneider A, Gonnet G. 2007. A phylogenomic study of human, dog, and mouse. PLoS Comput Biol. 3:e2.

Caron H, van Schaik B, van der Mee M, et al. (31 co-authors). 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, Campbell MJ, Winzeler EA, et al. (11 co-authors). 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell. 2:65–73.

Coghlan A, Wolfe KH. 2000. Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. Yeast. 16:1131–1145.

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

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

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

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

Elo A, Lyznik A, Gonzalez DO, Kachman SD, Mackenzie SA. 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, Bouhassira EE, Martin DI, Fiering S. 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, Tolaini M, Corbella P, Mamalaki C, Parrington J, Fox M, Miliou A, Jones M, Kioussis D. 1996. Locus control region function and heterochromatin-induced position effect variegation. Science. 271:1123–1125.

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

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

GU Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. 18:609–613.

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

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

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

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

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

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

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

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

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

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

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

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

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

Lahn BT, Pearson NM, Jegalian K. 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, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. 13:875–882.

Lefai E, Fernandez-Moreno MA, Kaguni LS, Garesse R. 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, Blumenthal T, Hurst LD. 2003. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. 13:238–243.

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

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

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

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

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

Lindsay SJ, Khajavi M, Lupski JR, Hurles ME. 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, Audic S, Claverie JM. 2003. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. Genome Biol. 4:P1.

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

Milot E, Strouboulis J, Trimborn T, et al. (11 co-authors). 1996. Heterochromatin effects on the frequency and duration of LCR-mediated gene transcription. Cell. 87:105–114.

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

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

Nishihara H, Hasegawa M, Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proc Natl Acad Sci USA. 103:9929–9934.

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

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

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

Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. Genome Biol. 8:R233.

Ranz JM, Maurin D, Chan YS, von Grotthuss M, Hillier LW, Roote J, Ashburner M, Bergman CM. 2007. Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biol. 5:e152.

Richards S, Liu Y, Bettencourt BR, et al. (52 co-authors). 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15:1–18.

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

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

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

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

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

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

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

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

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 USA. 101: 6062–6067.

Teichmann SA, Veitia RA. 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, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH. 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, McCarrey JR, Yang F, Page DC. 2001. An abundance of X-linked genes expressed in spermatogonia. Nat Genet. 27:422–426.

Waterston RH, Lindblad-Toh K, Birney E, et al. (223 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520–562.
Whitehead A, Crawford DL. 2006. Neutral and adaptive variation in gene expression. Proc Natl Acad Sci USA. 103:5425–5430.
Xing Y, Ouyang Z, Kapur K, Scott MP, Wong WH. 2007. Assessing the conservation of mammalian gene expression using high-density exon arrays. Mol Biol Evol. 24:1283–1285.
Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS. 8:15–24.
Yanai I, Korbel JO, Boue S, McWeeney SK, Bork P, Lercher MJ. 2006. Similar gene expression profiles do not imply similar tissue functions. Trends Genet. 22:132–138.

Marta Wayne, Associate Editor

Accepted April 19, 2008