Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals

Marie Sémon[1] *and Laurent Duret*

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Claude Bernard Lyon 1, Villeurbanne, France

Gene order is not random with regard to gene expression in mammals: coexpressed genes, and in particular housekeeping genes, are clustered along chromosomes more often than expected by chance. To understand the origin of these clusters and to quantify the impact of this phenomenon on genome organization, we analyzed clusters of coexpressed genes in the human and mouse genomes. We show that neighboring genes experience continuous concerted expression changes during evolution, which leads to the formation of coexpressed gene clusters. The pattern of expression within these clusters evolves more slowly than the genomic average. Moreover, by studying gene order evolution, we show that some clusters are maintained by natural selection and, therefore, have a functional significance. However, we also demonstrate that some coexpressed gene clusters are the result of neutral coevolution effects, as illustrated by the clustering of genes escaping inactivation on the X chromosome. Moreover, we show that, although statistically significant, constraints on gene orders have a limited impact on mammalian genome organization, affecting only 3–5% of the pool of human and murine genes. It had been hypothesized that coexpressed gene clusters might correspond to large chromatin domains. In contradiction, we find that most of these clusters contain only 2 genes whose coexpression may be due to transcriptional read-through or the activity of bidirectional promoters.

Introduction

The sequencing of whole genomes and the availability of expression data sets have introduced a new point of view in biology—the study of genes in their genomic context. This has made it possible to consider the expression of individual genes as a function of the expression of their neighbors. In all eukaryotic genomes so far analyzed, gene order is not random with regard to gene expression. Instead, there is a tendency for coexpressed genes to cluster significantly more than expected under the null model of a genome with no relationship between gene order and gene expression. A significant number of clusters have been found in several organisms, including yeast, Drosophila, nematode, mouse, and human (Hurst et al. 2004). Early works in mammals suggested that the genes in the clusters are those that are expressed in a specific tissue or that are highly expressed (Caron et al. 2001; Versteeg et al. 2003). However, other studies showed that these apparent patterns were by-products of the clustering of broadly expressed genes (Lercher et al. 2002). These observations suggest that coexpressed gene clusters could be a strong architectural component of the human genome. These results are striking because they could profoundly change the long-held assumption that genes are randomly located in our genomes. However, to assess the biological significance of this finding, it is still necessary to quantify the proportion of the genome covered by these clusters.

The molecular mechanisms underpinning coexpression are still unknown as gene expression is regulated at a number of levels. Chromatin structure is known to control the expression of genes and is an obvious candidate for the simultaneous regulation of neighboring genes (Sproul et al. 2005). It has been proposed that the eukaryotic genome is compartmentalized into chromatin domains (Hurst et al. 2004). Inside these domains, the chromatin can be in open conformation (the genes have the potential to be expressed) or in closed conformation (the genes cannot be expressed). The location of these domains may vary among cell types. It is therefore possible that genes that must be expressed in the majority of tissues should cluster in the zones of the genome where chromatin is in open conformation in the majority of tissues. In the same way, genes that must be expressed in a particular tissue could be localized in domains where chromatin is in open conformation in this particular tissue. These chromatin domains may be large enough to contain several genes (Hurst et al. 2004).

Alternatively, it is possible that coexpressed gene clusters are mainly due to small-scale mechanisms, such as regulatory elements (promoters or enhancers) shared by a few neighboring genes. For example, cases of divergently transcribed genes sharing a promoter sequence have been described in humans, and 10–20% of human genes could belong to such gene pairs (Trinklein et al. 2004).

Recently, it has been shown that gene expression evolves very rapidly in mammals (Khaitovich, Weiss et al. 2004), whereas gene order is evolutionarily very stable (Hillier et al. 2004). These observations are difficult to reconcile with the formation and maintenance of coexpressed gene clusters during evolution. We can see 2 different possible explanations. First, it is possible that genes located in coexpressed clusters have a much lower rate of expression evolution than other genes. An alternative explanation is that coexpression clusters change continuously in the genome. In that case, we expect to see a concerted evolution of the pattern of expression of neighboring genes. Note that coevolution of expression of neighboring genes (either due to chromatin domains or due to common regulatory elements) need not necessarily be of functional significance as genes might be switched on just because of their proximity to active genes (Spellman and Rubin 2002). In that case, mRNAs are not necessarily functional in the tissue, and this process could therefore be considered as neutral. We will refer to these nonfunctional coexpression clusters as "neutral coexpression cluster," as opposed to "functional coexpression clusters" that are composed of genes whose clustering is maintained by natural selection.

[1] Present address: Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland.

Key words: cluster, coexpressed genes, chromatin, gene order.

E-mail: semonm@tcd.ie.

Mol. Biol. Evol. 23(9):1715–1723. 2006
doi:10.1093/molbev/msl034
Advance Access publication June 6, 2006

To examine the significance of and evolutionary forces behind coexpressed gene clusters, we first quantified the extent of clusters using whole-genome expression data in human and mouse. Second, we studied the changes in expression between human and mouse to understand the processes of formation and maintenance of coexpressed gene clusters. Third, we assessed the evolutionary significance of coexpressed gene clusters, by trying to detect whether selection could be responsible for their maintenance.

Materials and Methods

Genome Data and Orthology Data

Human, mouse, and chicken gene sequences and locations were extracted from Ensembl (Birney et al. 2004) human genome (release 16.3, August 2003), mouse genome (release 18.33, November 2003), and chicken genome (release 24, October 2004).

Chicken protein sequences were extracted and compared with human proteins using BlastP (Altschul et al. 1997) to determine 11,192 pairs of orthologs by reciprocal best hit. Among these pairs, 5,763 correspond to human genes for which expression data can be computed using serial analysis of gene expression (SAGE) and expressed sequenced tag (EST) data (see below).

We used TreePattern to search in Homolens database, the gene families for which the tree topology matches a tree pattern corresponding to an 1:1 orthology relationship between human and mouse (Dufayard et al. 2005). We obtained 10,746 pairs of orthologous genes, for which expression pattern was inferred from EST data, in 17 tissues available in both species. To measure the rate of evolution of gene order, we computed the frequency of genes for which the 2 nearest neighbor genes are the same in both species: for each human gene (*Bh*), we considered its 2 flanking neighbors (*Ah* and *Ch*) among the data set of 10,746 genes in mouse, and we determined whether its ortholog in mouse (*Bm*) was flanked by *Am* and *Cm* (the orthologs of genes *Ah* and *Ch*).

Expression Data

We selected from GenBank (Benson et al. 2004; release 133, December 2002) 4,906,743 ESTs from human tissues and 3,660,463 ESTs from mouse tissues. cDNA libraries from cell culture, tumors, pooled organs, or unidentified tissues were excluded. To limit stochastic variations in expression measures, we only retained cDNA libraries that had been sampled with at least 10,000 ESTs. We retained 44 nontumoral tissues for human and mouse data sets. Gene-coding sequences (CDSs) were then compared with the EST data set by using MEGABLAST (Zhang et al. 2000). MEGABLAST alignments showing at least 95% identity over 100 nt or more were counted as a sequence match. This criterion was chosen to be low enough to allow the detection of most ESTs despite sequencing error but stringent enough to distinguish in most cases different members of highly conserved gene families. After adding all counts for libraries representing the same tissue type, we converted absolute EST counts to relative EST count (count per million). When there were several alternative-splicing variants, we randomly selected one CDS per gene. The final data set contains 19,482 human genes and 24,928 mouse genes.

SAGE experiment results were obtained on the SAGE Genie Web site (ftp://cgap.ncbi.nih.gov/Download; Liang 2002) for human data and on Gene Expression Omnibus site (http://www.ncbi.nlm.nih.gov/geo/; Edgar et al. 2002) for mouse data. We retained 141 libraries for the human data set (41 for mouse) containing more than 20,000 tags and not corresponding to tumoral tissues. The libraries were then grouped into 17 tissue types (11 for mouse).

To determine the expression pattern of a given gene with SAGE, it is necessary to know the sequence of its 3′ end (3′ untranslated region). Given the inaccuracy of gene prediction methods, we decided to restrict our analyses to genes for which an mRNA sequence was described and manually curated in the RefSeq database (Pruitt et al. 2003). The tag (10 bp upstream of the most 3′ *Nla*III restriction site) was extracted from the RNAs. In some cases, one tag may match to more than one Refseq mRNA. We looked at the genomic location of these mRNAs to determine whether they correspond to alternative transcripts of a same gene or to different genes. In the latter case, tags ambiguously located were removed from the data set. When there were several alternative-splicing variants, we randomly selected one RNA per gene. Normalization of the absolute tag count was done as described for EST data. Association between Refseq RNAs and Ensembl genes permitted to get the location of the RNAs. We retained 13,435 human and 8,951 mouse RefSeq RNAs that are nonredundant and unambiguously located, respectively, on the human and mouse genome.

All the analyses were done using qualitative expression data, either the presence or absence for each tissue or expression breadth (the number of tissues where a gene is expressed). After removing tandem duplicates (see below), we retained human genes for which EST and SAGE data are available for 14 tissues common to both methods and those that are expressed in at least one of these 14 tissues for both methods. We finally obtained a data set composed of 9,765 human genes.

Cluster Identification

Gene coexpression was estimated by the index of common expression ($\mathrm{ICE}_{a,b}$) for each gene pair *a*, *b* (Lercher et al. 2002):

$$\mathrm{ICE}_{a,b} = \frac{\sum_t f_{a,t} f_{b,t}}{\sqrt{(\sum_t f_{a,t} \sum_t f_{b,t})}},$$

where t runs over all the tissues and $f_{a,t}$ indicates whether the gene a is expressed in the tissue t ($f_{a,t} = 1$) or not ($f_{a,t} = 0$).

A gene cluster is a contiguous group of coexpressed genes (the index is higher than 0.5 for all gene pairs in the cluster). Following previous report (Lercher et al. 2002), tandem duplicated genes separated by less than 1 Mb were detected in the data set, using a conservative criterion (BlastP *E*-value, $E < 0.2$). One of the 2 duplicates was picked randomly and removed from the data set. Genes were ordered according to their position on the chromosomes.

Evaluation of Coevolution

To evaluate coevolution phenomenon, we regarded as neighbors 2 genes that are adjacent on chromosomes of human and mouse and at a distance of less than 1 Mb. We used a data set containing 8,488 pairs of genes that have not undergone a rearrangement since the divergence between human and mouse. We built a simple model to evaluate quantitatively the coevolution of gene expression in neighboring genes.

We consider the evolution of the expression of 2 adjacent genes between human and mouse. As we only consider 2 sequences, the ancestral state is not known. However, the model is reversible; it is thus equivalent to model the evolution of expression from human toward mouse or the reverse. The expression in mouse will be arbitrarily considered to be ancestral. Each gene can either be expressed (denoted 1) or not expressed (0) in each tissue. The model uses 3 parameters. We define p as the probability that a gene is not expressed in mouse and becomes expressed in human and q as the probability that a gene is expressed in mouse and not in human. r is the probability of coevolution (unknown), that is, the probability that the change of expression of a gene is propagated to its neighbor. We assume that p, q, and r are constant along the sequence and during evolution. The theoretical frequencies f can be expressed according to these probabilities. For example, $f1$ is the fraction of the adjacent genes that were not expressed in mouse and are still not expressed in human. These genes have not changed their pattern of expression (multiple changes are not taken into account in this very simple model): $f1 = (1 - p)^2$. $f2$ represents the fraction of the adjacent gene pairs that are not expressed in mouse but are in human. In this case, it is possible that each gene underwent an independent change of expression (p^2). It is possible also that the expression pattern of one gene changed and that this change propagated to the second gene ($2pr$). Therefore, one can write $f2 = p^2 + 2pr$. The theoretical frequencies are entered into a matrix according to this simple principle (Figure 1 in Supplementary Material online).

The estimation of the parameters p, q, and r can be carried out by minimizing an adjustment criterion (*nlm* under *R*). In entry, it needs a rough estimate of the parameters (pguess). From "pguess," the parameters are optimized by iteration. For this reason, it is possible that the parameters obtained correspond, in fact, to a local minimum of this function. It is thus necessary to test if the estimate of the parameters is dependent on the initial conditions. We tested the validity of the model by simulation. We start from an artificial chromosome of mouse containing 1,000 genes, of which 40% are expressed. The principle is to simulate the evolution of this "mouse chromosome" to obtain a "human chromosome." The evolution is simulated while making the expression of genes change a certain number of times: the parameters *pt*, *qt*, and *rt* (true values) are given. By comparing the mouse chromosome and human chromosome, we built the matrix of the observed changes of expression and extracted from it *pe*, *qe*, and *re* (estimated values of the parameters). We compare finally the values *pe*, *qe*, and *re* with the values *pt*, *qt*, and *rt* to check that our model estimates the good values of the parameters. We first inserted changes of expression in the sequence of mouse, with $pt = 0.02$ and $qt = 0.01$ without coevolution ($rt = 0$). We found good estimates of the probabilities: $pe = 0.0194$, $qe = 0.01$, and $re = 0.00$. To check that optimization does not depend on the initial conditions, we varied pguess, and we obtain the same results. Then, we carried out various simulations for values of *rt* varying from 0 to 0.12 (with $pt = 0.02$ and $qt = 0.01$, Figure 2 in Supplementary Material online). The estimated value of coevolution *re* is very close to *rt*, the real coevolution (correlation $R^2 = 0.99$). Finally, we inserted changes of expression in the sequence of mouse, with $pt = 0.2$ and $qt = 0.1$. We carried out various simulations for values of *rt* varying from 0 to 0.12. The probabilities *pe* and *qe* are underestimations of *pt* and *qt*. This is due to frequent multiple changes of expression ($pe = 0.159$, $qe = 0.059$). The estimated coevolution *re* is not very close to *rt*, the real coevolution. More exactly, the correlation is very good between 2 measures ($R^2 = 0.99$), but *re* underestimates the actual value of the coevolution (straight regression line $rt = 1.459$; $re = 0.0015$). These simulations show that the estimate of the coevolution is reliable when the number of changes is small. On the other hand, when it is too high, the method underestimates the actual value of the coevolution. On average, 30% of genes changed expression in a given tissue between human and mouse (median: 26%). This implies that the coevolution is likely to be difficult to estimate correctly (because p and q are high). The application of the model to the 30% of the tissues for which the number of changes of the expression between human and mouse is weakest gives an average coevolution value of 0.025. This probability is in worst case an undervaluation of the true value. We thus estimate that the change of expression of a gene has a 2.5% chance to propagate to its neighbor.

Evaluation of the Impact of Coevolution of Expression

Does coevolution have an impact on the formation of coexpressed gene clusters? To answer this question, we carried out simulations: we initially broke the gene clusters by permuting gene order. Then, expression changes are carried out on these genes in order to maintain the percentage of expressed genes in each tissue (e.g., if 20% of genes are expressed in this tissue, we choose $pt = 0.2$ and $qt = 0.8$ for this tissue). At each step of time, a gene is picked randomly to change expression, in a tissue randomly chosen. The change of expression can or cannot be propagated to close genes with a probability *rt*. Simulations are stopped when the number of genes differentially expressed is comparable to the average number observed between human and mouse (30%).

Genes Escaping Inactivation on the X Chromosome

We used the data set published by Carrel and Willard (2005) that provides an X-inactivation profile of the human Xi (X inactivated) chromosome in 9 different individuals. We retained 619 genes that are not located on the PAR region. We considered as expressed the genes escaping Xi inactivation in all or all but one essayed individuals.

Results and Discussion

Quantification and Description of Coexpressed Gene Clusters

To study the impact of coexpressed gene clusters on human genome organization, we first estimated the percentage of genes that belong to a cluster. We selected a data set of 9,765 human genes for which the pattern of expression (i.e., the list of tissues where genes are expressed) could be estimated using SAGE data in 14 normal tissues (see Materials and Methods). As an independent estimation, we also obtained the pattern of expression of these genes in these 14 tissues using EST data. Among these genes, we identified clusters of coexpressed genes in the human genome, using the method published previously by Lercher et al. (2002; clusters were determined independently with EST and SAGE data). This method consists in grouping adjacent genes that share more than 50% of the tissues in which they are expressed (see Materials and Methods). Note that genes duplicated in tandem may have retained similar expression profiles simply because of their common origin (Lercher et al. 2002). To exclude such trivial clusters, only one gene from each set of tandem paralogs was retained in the data set. Using these criteria, we found that 65% of the human genes belong to coexpressed gene clusters. This result depends on the method used to detect clusters but suggests that clustering of coexpressed genes is a genome-wide phenomenon. The distribution of cluster size in a number of genes is highly biased toward small clusters (fig. 1).

Even under a null model of gene clustering, with no relationship between gene order and expression, we expect some coexpressed genes to be clustered by chance. For instance, 2 randomly chosen genes that are broadly expressed are likely to be coexpressed according to the method we used to measure coexpression. We determined whether the number of clusters observed in the data was significantly greater than that expected under such a model. We therefore estimated the number of clusters corresponding to background noise. We created 1,000 randomly permutated genomes by independently permuting gene order and calculated the average number of clusters for these simulated genomes. The number of solitary genes (that do not belong to a cluster) is significantly lower in the real genome than in permutated genomes, for both SAGE and EST data (fig. 1, $P < 10^{-16}$). This test is in agreement with previous results (Lercher et al. 2003) and shows that there is a significant clustering of coexpressed genes in the human genome. This phenomenon is not very important quantitatively: as shown in figure 1, the number of coexpressed gene clusters obtained after permutations is close to the real number of clusters. In our simulated data set, an average of 60% of the genes occur in clusters (for EST data). In the real data, this figure is 65%. The difference represents 5% (for EST) and 3% (for SAGE) of the total number of genes. Using the same cluster detection method, we found a significant clustering in the mouse genome (EST and SAGE). This confirms previous results obtained on a small data set (Williams and Hurst 2002). The excess of mouse genes belonging to a cluster in the observed data set compared with the random expectations is comparable to what we observed in human (5% and 3% of the data sets for EST and SAGE, respectively). Thus, the clustering of coexpressed genes involves only a tiny fraction of the gene repertoire in mammals beyond random expectation.

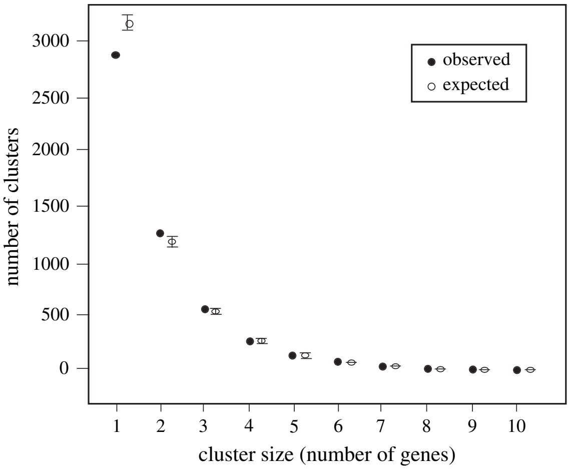

FIG. 1.—Frequency distribution of the size (in number of genes) of clusters of coexpressed genes in the human genome. The data set includes 9,765 genes for which expression data were obtained using 14 tissues from SAGE libraries. We used the method published by Lercher et al. (2002) to detect coexpressed gene clusters in this data set and in data sets obtained after randomly permuting gene positions. Filled circles represent the number of observed clusters for each cluster size (in number of genes). Open circles represent the number of clusters for each size of cluster averaged in 1,000 independent permutations of gene positions. Error bars represent confidence intervals ($P = 5\%$) of the means. As it was previously shown (Lercher et al. 2002), there is a significant clustering in the human genome. Indeed, the observed number of singleton genes (that do not belong to a cluster) is significantly lower ($P < 10^{-16}$) in the real genome than in permutated genomes. However, the observed number of clusters is close to the expected number.

Formation and Maintenance of Coexpressed Gene Clusters

Genomic rearrangements can be responsible for the formation of coexpression clusters, as it was shown for the DAL cluster in yeast (Wong and Wolfe 2005). However, because gene order is evolutionarily very stable in mammals (Hillier et al. 2004) and because gene expression evolves by contrast very rapidly (Khaitovich, Weiss et al. 2004), it is difficult to understand how coexpressed gene clusters are maintained during evolution. It is of course possible that the evolution of expression does not occur at similar rates across the genome because genes belonging to coexpressed gene clusters have a particularly well-conserved pattern of expression. It is also possible that neighboring genes experience concerted expression changes during evolution that could both create and maintain coexpression clusters. We refer to this process as coevolution of patterns of expression. Studying the evolution of the pattern of expression within the clusters is therefore helpful to understand how they are created and maintained.

Coevolution of the Patterns of Expression of Neighboring Genes

We will first focus on the coevolution of patterns of expression of neighboring genes. In cases of coevolution,

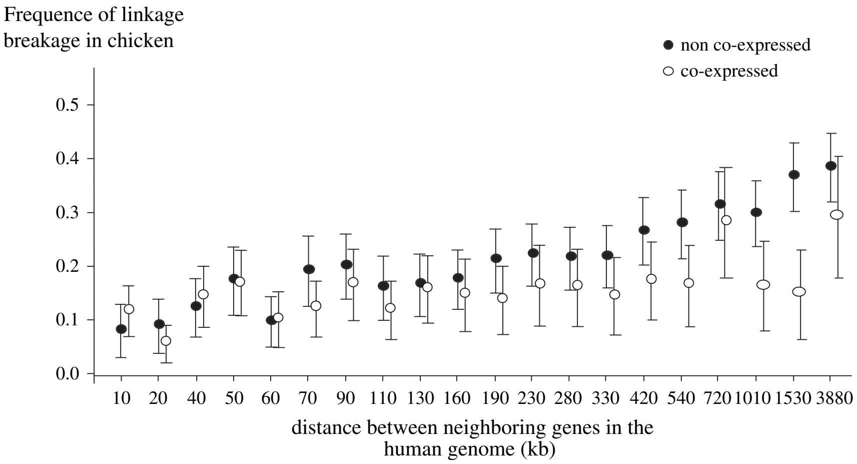

FIG. 2.—Conservation of the linkage between human and chicken according to the physical distance between genes in the human genome, for genes that are coexpressed or not coexpressed in humans. The data set includes 5,763 genes for which orthologs (determined using reciprocal best hits) are known in humans and chickens and that are expressed in at least in one tissue in humans. The expression pattern of human genes was estimated with SAGE data. For each pair of neighbor genes in the human genome, we determined 1) whether they are coexpressed (they belong to the same cluster) or not in humans, 2) the physical distance between the 2 genes in the human genome, and 3) whether their orthologs are linked in the chicken genome. The relationship between the frequency of linkage conservation and intergenic distance is indicated for coexpressed genes (filled circles) and non–coexpressed genes (open circles). In both cases, the frequency of linkage conservation decreases with increasing intergenic distance ($P = 10^{-7}$ for coexpressed genes and $P < 10^{-16}$ for non–coexpressed genes). For a given intergenic distance, there is a significant difference in the frequency of linkage conservation between coexpressed and non–coexpressed genes ($P = 10^{-6}$).

if the expression pattern of a gene is modified so that it is now expressed in a particular tissue, its neighboring genes have a chance to become expressed too. Conversely, when a gene ceases to be expressed in a tissue, its neighbors could also cease to be expressed. Coevolution of expression can therefore be studied by looking at the consequences of the expression change of a gene on its neighbors during evolution. We compared the changes of gene expression between human and mouse. If coevolution occurs, changes of expression occurring in the same direction in 2 close genes will be more frequent than expected (fig. 3*a*), whereas changes of expression in opposite directions will be less frequent than expected (fig. 3*b*).

To test for an excess of coordinated changes of expression pattern, we studied differences in expression between human and mouse. Expression pattern evolves very quickly in mammals (Makova and Li 2003; Khaitovich, Muetzel et al. 2004; Yanai et al. 2004), and it is possible that several changes of the expression pattern for any gene have taken place since the human/mouse divergence. Despite this divergence, we choose these species because human and mouse are the closest species for which sufficient EST data are available (6.1 million ESTs for human, 4.3 million ESTs for mouse in March 2005). We collected a data set of 8,488 orthologous human/mouse gene pairs for which EST data are available in 17 tissues in both species. We then counted the number of coordinated changes in expression of neighboring genes and compared this with the number of such changes expected by chance.

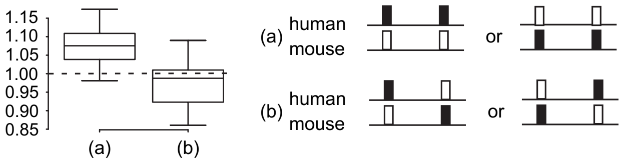

FIG. 3.—Consequences of the coevolution of expression. (*a*) Top: various possibilities for 2 couples of orthologous genes that are adjacent in the human genome and in the mouse genome. Both ortholog couples have changed expression in a tissue given between humans and mice. (*b*) Bottom: a representation of the distribution of the ratios of the observed on the expected expression changes in the same direction (*Rs*) and in opposite directions (*Ro*). *Rs* is significantly different from 1 ($P = 10^{-4}$), and *Ro* is marginally significant ($P = 0.06$, Wilcoxon tests).

Under the assumption that a change in the expression of a gene does not influence the expression of its neighbor and knowing the number of genes that have changed in expression since the divergence of human and mouse, it is possible to calculate the expected number of changes in expression (without coevolution) of 2 adjacent genes (see Materials and Methods). This calculation corrects for a possible difference of sampling in the expression libraries from the 2 species. It is thus possible to calculate in each tissue the average observed and expected numbers of expression changes for 2 adjacent genes in the same direction and in opposite directions and then the ratios of observed to expected changes in the same direction (*Rs*) and in opposite directions (*Ro*). If there is a covariation of the changes of expression for close genes, one expects $Rs > 1$ and $Ro < 1$. These ratios are presented in figure 3 (see also table 1 in Supplementary Material online). *Rs* is significantly different from 1 (average on the 17 tissues: $Rs = 1.07$, $Rs > 1$ with $P = 10^{-4}$; $Ro = 0.98$ on average, $Ro < 1$ with $P = 0.06$, Wilcoxon tests), indicating that there is a coevolution of the expression of neighbor genes.

Coevolution of expression depends on the distance between the genes. When only the 2,619 pairs of genes more distant than 50,000 bp are retained (30% of the data), *Rs* is no longer significantly different from 1 ($Rs = 1.01$ on average, $P = 0.5$). On the other hand, when the 2,921 pairs of genes closer than 50,000 bp are retained (30% of the data), there is a significant coevolution ($Rs = 1.07$ on average, $N = 2{,}921$ pairs, $Rs > 1$ with $P = 0.004$). Coevolution also depends on the direction of transcription. Coevolution is observed when the genes are transcribed in the same direction (4,343 pairs of genes, $Rs = 1.09$ on average, $Rs > 1$ with $P = 0.001$) and when the genes are transcribed in divergent directions ($Rs = 1.11$ on average, $Rs > 1$ with $P = 0.003$, 1,240 pairs of genes) but not for genes transcribed in convergent directions ($Rs = 1.02$ on average, $P = 0.25$, 1,241 pairs of genes).

Note that the ratios *Rs* are significant even though the number of points (17 tissues) is rather small. It is also striking that the distance between the genes in convergent orientation is smaller than for the pairs in the same or in divergent orientation (115,500, 131,100, and 142,900 bp on average for the pairs in convergent, same, and divergent orientations, respectively, all of the pairwise comparisons with $P < 10^{-16}$, Wilcoxon tests), but there is nonetheless no coevolution for the pairs in convergent orientation. We propose that most of these clusters areeither the result of transcription read-through phenomenon or of the presence of bidirectional promoters.

We built a simple model to quantify coevolution (see Materials and Methods) and estimate that the change of expression of a gene has 2.5% of chances of propagating to its neighbor. This demonstrates that coevolution of expression is a significant phenomenon in human genome, even if it is a quantitatively weak effect. What is its impact on the formation of coexpressed gene clusters? To answer this question, we performed simulations (see Materials and Methods). When using the coevolution value we had estimated, the number of solitary genes obtained after simulations is not significantly different from the observed number. This suggests that coevolution of expression is a sufficient phenomenon to create coexpression clusters (Figure 3 in Supplementary Material online).

Slower Evolution of the Expression in the Coexpressed Gene Clusters

Slower evolution of expression in coexpressed gene clusters could also resolve the apparent contradiction between evolutionary rates of change in expression and gene order. To test this hypothesis, we retained 8,948 orthologous gene pairs between human and mouse that are expressed in at least one tissue in each species out of a subset of 16 tissues common to both species. We calculated the coexpressed gene clusters in each species and found that genes located in a cluster in one species tend to be located in a cluster in the other species more often than expected (chi-square test, $P < 10^{-16}$). We found also that the pattern of expression of genes located in clusters is more conserved than expected between human and mouse. (Average frequency of tissues where the genes are expressed in human and in mouse: 46% for genes located in clusters in human, 35% for genes outside clusters, Wilcoxon test: $P < 10^{-16}$.)

However, this pattern could be due to the excess of housekeeping genes in coexpressed gene clusters: genes belonging to a cluster in human are expressed in more tissues than others (on average, genes belonging to a cluster are expressed in 7.6 tissues in human vs. 4.0 for the other group of genes, $P < 10^{-16}$). Both the sequences and the patterns of expression of housekeeping genes tend to be slowly evolving because of higher selective pressure (Khaitovich et al. 2005). Therefore, we only retained in our data set the top 30% of genes with the widest expression breadth (more than 9 tissues, 1,930 genes). The expression breadth of clustered genes is no longer significantly different than that of nonclustered genes (Wilcoxon test, $P = 0.1$), but the pattern of expression is still more conserved in clusters ($P = 0.007$, Wilcoxon test). We conclude therefore that the expression patterns of genes belonging to coexpressed gene clusters tend to evolve at a slower rate.

It therefore appears that both coevolution of expression and an excessive conservation of the pattern of expression inside coexpressed gene clusters make it possible to create and to maintain coexpressed gene clusters in the human genome.

Evolutionary Significance of Coexpressed Gene Clusters

Coevolution of the pattern of expression in neighboring genes seems to create clusters of coexpression. These observations are important because they show how coexpressed gene clusters are created. However, the functional significance of these clusters remains to be determined. Clusters could be of functional relevance and therefore maintained by selective pressure. Alternatively, they could be nonfunctional and maintained by a neutral phenomenon of coevolution of expression.

Some Coexpressed Gene Clusters Are Functional

If clusters are functional, they should be more conserved than expected during evolution. In other words, the probability of linkage retention after a long time of divergence should be higher for genes that belong to the same coexpression cluster than for genes that are not coexpressed.

To test this hypothesis, we analyzed the conservation of linkage between mammals and birds, using 2 species, human and chicken (*Gallus gallus*), for which the complete genomes are now available (Lander et al. 2001; Hillier et al. 2004). We choose to compare these 2 species because their last common ancestor is ancient enough (310 MYA; Hillier et al. 2004) to ensure a certain number of rearrangements in the human lineage since this divergence. Each human gene was associated by reciprocal best hit to its ortholog in the chicken genome. The resulting data set contains 5,763 genes expressed in at least one tissue in human and with an ortholog identified in chicken. We obtained the same results by computing orthologs using a phylogenetic approach (data not shown). We checked for each pair of adjacent genes in the human genome whether the corresponding orthologous genes were linked or not in the chicken genome. We define as a case of conserved linkage pairs of genes whose orthologous genes in the chicken genome are either adjacent or have at most one intermediate gene in the chicken genome. All the other cases were defined as linkage breakage. Cases of linkage breakage can thus correspond to genes located on different chromosomes in chicken or to genes located in distant regions of the same chromosome. We find that the frequency of linkage breakage in chicken is lower for genes that belong to the same coexpression cluster in human (18%) than for other genes (25%). This observation is consistent with the hypothesis that there is a selective pressure to maintain linkage between coexpressed genes. However, interpreting this result is not trivial because coexpressed genes tend to cluster into gene-dense regions (Versteeg et al. 2003): for SAGE data, the average intergenic distance between 2 neighboring genes that are coexpressed in humans is 284 kb (median = 89 kb) as compared with 617 kb (median = 222 kb) for 2 neighboring genes that are not coexpressed (Wilcoxon test, $P = 10^{-16}$; the mean values are 281 and 584 kb, respectively, for the EST data set). Thus, the higher linkage conservation of coexpressed genes could be simply a consequence of smaller intergenic distances. To test this hypothesis, we analyzed the frequency of linkage conservation according to the length of the intergenic spacer for pairs of genes that are coexpressed or not. As shown in figure 2 for SAGE data, linkage breakage is lower for genes that are coexpressed in humans even when intergenic distance is taken into account ($P < 10^{-16}$, generalized linear model, fig. 2). The same results are obtained with EST data ($P < 10^{-16}$). In other words, pairs of coexpressed genes tend to occur in gene-dense regions and show a stronger conservation of linkage than other genes (not coexpressed) located in regions of similar gene density. Thus, at the evolutionary scale considered here, we found evidence of a selective pressure to maintain linkage between coexpressed genes in at least some of the coexpressed gene clusters.

This finding is in agreement with Singer et al. (2005) who found that linkage conservation between human and mouse is higher for genes that are coexpressed in human.

We checked whether linkage conservation depends also on their relative orientation. We first focused on pairs of genes transcribed in the same orientation in human (N = 1,895). In this subset, we found that linkage conservation is higher for gene pairs belonging to the same cluster in human even when intergenic distance is taken into account (EST: $P = 0.02$; SAGE: $P = 10^{-5}$). The test is also significant for pairs of genes in divergent orientation ($N = 991$, EST: $P = 0.002$, SAGE: $P = 0.04$) but not for pairs of genes in convergent orientation ($N = 998$, EST: $P = 0.4$; SAGE: $P = 0.07$).

Some Coexpressed Gene Clusters Are Not Functional

We have shown that some clusters are functional because they are maintained more than expected during the course of evolution. This does not exclude that some clusters may result from nonfunctional coevolution of expression. A recent study of the X chromosomes in women has shed light on the existence of such a neutral phenomenon (Carrel and Willard 2005). In female mammals, one of the X chromosomes is inactivated to compensate for the difference in gene dosage with XY males. Consequently, most genes on this copy are silenced. Some genes escape this silencing and are then expressed from both active (Xa) and inactive X (Xi) chromosomes. A recent study has shown that at least 16% of genes escape inactivation on Xi, and an additional 10% of genes are variably inactivated among individuals. The majority of the genes (63%) having a homolog on the Y chromosome escape inactivation. This was expected because there is no need for dosage compensation for genes that are expressed both from the X and the Y (Charlesworth 1998). However, many of the genes that escape from inactivation do not have a homolog on the Y chromosome. Because genes that escape inactivation tend to be clustered, Carrel and Willard (2005) have suggested that they may "lie within an epigenetic domain containing at least one X-linked gene with a Y homologue." We quantified the scale of this phenomenon by reanalyzing this data set. We split the data set (619 genes located on the X chromosome, among which 74 escape inactivation, see Materials and Methods) into genes that have a homolog on the Y (35 "Y-genes") and genes that do not (584 non–Y-genes). For each Y-gene, we attributed a set of neighboring non–Y-genes that were closer to this Y gene than any other. We then plotted the frequency of escaping inactivation of the neighboring genes as a function of their distance from their Y-gene. The probability of escaping inactivation depends strongly on this distance (chi-square test, $P = 10^{-15}$). For instance, it is 2 times higher than expected for genes adjacent to a Y-gene (fig. 4). This probability decreases sharply as the distance to the Y-gene increases, indicating that this effect acts at a very small scale (1–2 genes). This is comparable to the size of the clusters of coexpressed genes in the whole genome (fig. 1), even though 2 intervening genes appear still coexpressed on the X (fig. 4). This longer range of coexpression on the X chromosome may be due to the fact that our data on the X chromosome are both more precise and more complete than on the autosomes, permitting a better evaluation of the phenomenon. To investigate further the mechanisms creating coevolution on the X chromosome, we have represented the coexpression on the X chromosome depending on the relative orientation of a non–Y-gene and the nearest Y-gene (same/divergent/convergent). As in the autosomal data set, we observe coexpression for genes that are in the same orientation than the nearest Y-gene, and for genes in divergent orientation, but not for genes in convergent orientation (Figures 4–6 in Supplementary Material online).

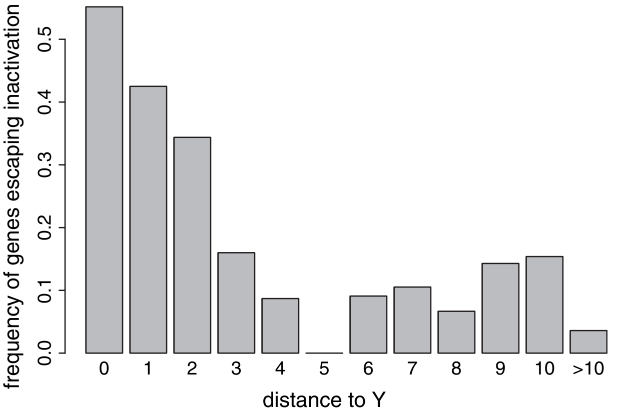

FIG. 4.—Frequency of gene expression on the Xi chromosome function of the distance (in number of genes) to the nearest gene having an homolog on the Y chromosome (data set from Carrel and Willard 2005). The genes having a homolog on the Y chromosome ($N = 29$, distance = 0) have a much higher probability to be expressed on the Xi chromosome ($P = 0.6$) than the genes without homolog ($N = 577$, $P = 0.1$). However, this probability increases for genes located in proximity to the genes having a homolog on the Y chromosome (3–4 times higher than average for the closest neighbors, distance 1 or 2).

It is easy to propose an explanation why Y-genes, which are present in 2 copies in males and females, escape inactivation. Many of the X genes that have retained a homolog on the Y probably correspond to non–sex-specific genes that have to be expressed at a high level and, hence, that have to be present in 2 copies both in males and females. Thus, this selective pressure for high expression may explain why most of these Y-genes escape X inactivation in females. But how to explain that genes located nearby Y-genes tend to escape inactivation? Some non–Y-genes may escape inactivation because they need to be expressed in higher level in females than in males. One might imagine that these genes have been translocated into the immediate vicinity of Y-genes so that they can escape inactivation. To test this hypothesis, we investigated the location of these genes in the last common ancestor of mammals and birds, that is, before the differentiation of the X and Y sex chromosomes. We determined the position in the chicken genome of each pair of genes consisting of a non–Y-gene escaping inactivation (in at least one individual) and its nearest Y-gene neighbor. We restricted the data set to the pairs where the non–Y-gene is in the immediate vicinity of the Y-gene (less than 10 genes) because the coevolution of expression acts at a very small scale. We found that in 94.4% of the cases (17 cases out of 18), the linkage between the non–Y-genes escaping inactivation and the nearest Y-gene predates the human/chicken divergence. Hence, the linkage between these genes cannot be explained by differences in dosage constraints between males and females because it predates the differentiation of X and Y. We therefore conclude that the only reason for these non–Y-genes to escape inactivation is that they are located near a Y-gene. The expression in double dosage of these genes is probably tolerated because it has no deleterious impact on the phenotype and is simply a nonfunctional consequence of the expression of the Y-gene.

Conclusion

It has been suggested that clusters of coexpressed genes are a significant organizational and even functional component of the architecture of our genome. Here we show that there is a significant clustering of coexpressed genes in human and mouse genomes, but this is a very weak effect: the number of clusters observed in these genomes only slightly exceeds the number expected by chance. This excess corresponds to only 3–5% of mammalian genes. This result depends on the quality of the expression data set and on the definition of coexpression clusters. However, we think that this estimation is reliable because our results are consistent with those obtained by Versteeg et al. (2003) who found 30 clusters of highly expressed genes (RIDGE) representing 1,359 transcription units in the human genome. They are also in agreement with Megy et al. (2003), who detected 31 clusters distributed in 3 human chromosomes (20, 21, and 22) and totalize 64 genes. A recent study also shows that 9% of the genes belong to coexpressed gene clusters defined using microarray data (Liu et al. 2005). Therefore, we can state that the clustering of coexpressed genes is an exception and not a rule in our genome and cannot therefore impact considerably on the genome structure.

The clusters that are found are highly dependent on the method used to evaluate coexpression. A few very large coexpression clusters have been reported (see, for instance, Lercher et al. 2002; Versteeg et al. 2003). Our results do not contradict these observations, but we show that these large clusters are in minority in the human genome.

To understand the processes of formation and maintenance of coexpressed gene clusters, we studied the changes in expression between human and mouse. We validate the 2 hypotheses that we proposed in introduction to understand the presence of coexpression cluster in the context of a rapid evolution of expression pattern in mammals. We show that neighboring genes experience concerted expression changes during evolution. This phenomenon of coevolution is weak quantitatively but can nonetheless generate clusters of coexpression in the genome. The change of the pattern of expression is on average rapid during the course of evolution genes, and, by contrast, the expression of genes that belong to a coexpression cluster evolves slower.

We assessed the evolutionary significance of coexpressed gene clusters. We observe that coexpressed gene clusters are maintained more often than expected by chance during evolution. We conclude therefore that some clusters have a functional significance. By contrast, another population of coexpressed gene clusters is probably due to neutral coevolution effects, as illustrated by the clustering of genes that escape inactivation on the X chromosome.

The molecular mechanisms involved in coexpression of neighbor genes are still unknown. It had been hypothesized that chromatin domains containing several genes play a role in the creation of coexpressed gene clusters (Hurst et al. 2004; Sproul et al. 2005). Alternatively, coexpressed gene clusters could be mainly maintained by small-scale mechanisms, such as regulatory elements (promoters or enhancers) shared by a few neighboring genes. We studied the size of the clusters and found that most of them contain 2–3 genes. The small-scale effects observed are compatible with our study of coevolution of expression showing no particular coevolution of expression for the most distant pairs of genes. We also show that the phenomenon of genes escaping inactivation on the Xi chromosome acts at a very small scale (1–2 genes). The phenomenon generating the clusters act not only at a small scale but also for pairs of genes in a peculiar orientation: coevolution of the patterns of expression and conservation of the clusters of coexpressed genes is only visible for pairs of transcripts in divergent or in the same orientation. Given the size of clusters we observe and the orientation of the genes located inside these clusters, we propose that most of these clusters are either the result of transcription read-through phenomenon or of the presence of bidirectional promoters.

Supplementary Material

Table 1 and figures 1–6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Centre National de la Recherche Scientifique. We would like to thank Meg Woolfit for critical reading of the manuscript and 2 reviewers for their helpful comments.

Literature Cited

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2004. GenBank: update. Nucleic Acids Res 32:D23–6.

Birney E, Andrews D, Bevan P, et al. (48 co-authors). 2004. Ensembl 2004. Nucleic Acids Res 32:D468–70.

Caron H, van Schaik B, van der Mee M, et al. (13 co-authors). 2001. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science 291:1289–92.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434:400–4.

Charlesworth B. 1998. Sex chromosomes: evolving dosage compensation. Curr Biol 8:R931–3.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics 21:2596–603.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–10.

Hillier LW, Miller W, Birney E, et al. (174 co-authors). 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.

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

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 309:1850–4.

Khaitovich P, Muetzel B, She X, et al. (14 co-authors). 2004. Regional patterns of gene expression in human and chimpanzee brains. Genome Res 14:1462–73.

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:E132.

Lander ES, Linton LM, Birren B, et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nat Genet 31:180–3.

Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. 2003. A unification of mosaic structures in the human genome. Hum Mol Genet 12:2411–5.

Liang P. 2002. SAGE Genie: a suite with panoramic view of gene expression. Proc Natl Acad Sci USA 99:11547–8.

Liu C, Ghosh S, Searls DB, Saunders AM, Cossman J, Roses AD. 2005. Clusters of adjacent and similarly expressed genes across normal human tissues complicate comparative transcriptomic discovery. OMICS 9:351–63.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13:1638–45.

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

Pruitt KD, Tatusova T, Maglott DR. 2003. NCBI Reference Sequence project: update and current status. Nucleic Acids Res 31:34–7.

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 22:767–75.

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

Sproul D, Gilbert N, Bickmore WA. 2005. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet 6:775–81.

Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. 2004. An abundance of bidirectional promoters in the human genome. Genome Res 14:62–6.

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.

Williams EJ, Hurst LD. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J Mol Evol 54:511–8.

Wong S, Wolfe KH. 2005. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. Nat Genet 37:777–82.

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.

Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–14.

Dan Graur, Associate Editor

Accepted June 1, 2006