Clusters of Co-expressed Genes in Mammalian Genomes Are Conserved by Natural Selection

Gregory A. C. Singer,[1] *Andrew T. Lloyd,*[2] *Lukasz B. Huminiecki,*[3] *and Kenneth H. Wolfe*

Department of Genetics, Smurfit Institute, University of Dublin, Trinity College, Dublin, Ireland

Genes that belong to the same functional pathways are often packaged into operons in prokaryotes. However, aside from examples in nematode genomes, this form of transcriptional regulation appears to be absent in eukaryotes. Nevertheless, a number of recent studies have shown that gene order in eukaryotic genomes is not completely random, and that genes with similar expression patterns tend to be clustered together. What remains unclear is whether co-expressed genes have been gathered together by natural selection to facilitate their regulation, or if the genes are co-expressed simply by virtue of their being close together in the genome. Here, we show that gene expression clusters tend to contain fewer chromosomal breakpoints between human and mouse than expected by chance, which indicates that they are being held together by natural selection. This conclusion applies to clusters defined on the basis of broad (housekeeping) expression, or on the basis of correlated transcription profiles across tissues. Contrary to previous reports, we find that genes with high expression are not clustered to a greater extent than expected by chance and are not conserved during evolution.

Introduction

Prokaryotes use a simple yet elegant system to regulate the expression of their genes. Genes belonging to the same functional pathways are often packaged into operons, which are transcribed into a single mRNA. Although this system works well in the prokaryotic context, operons appear to be very rare in eukaryotes and have only been discovered in a few organisms, most notably nematode worms (Zorio, et al. 1994; Blumenthal, 1998; Blumenthal et al. 2002) where it is estimated that 15% of genes within *Caenorhabditis elegans* are contained in operons. However, the mechanisms involved in the processing of polycistronic mRNAs are quite distinct in *C. elegans* compared to bacterial genomes, so it is likely that nematode operons are independent innovations within their lineage (Blumenthal et al. 2002). Despite the absence of operons, eukaryotes are still capable of a very fine level of control over gene transcription. However, this is accomplished through the use of *trans*-acting factors that do not require the co-transcribed genes to be in close proximity to each other (Niehrs and Pollet, 1999). Does this mean that the order of genes in the eukaryotic genome is random? Certainly, if the positioning of genes within the genome is not important to transcriptional regulation then the high rate of genome rearrangement events in eukaryotic genomes will lead to the complete randomization of gene order in a short period of time (Huynen, Snel, and Bork, 2001). A number of studies, however, indicate that there is some gene organization in eukaryotic genomes, and that *cis*-acting regulatory factors may play a larger role than previously thought (Hurst, Pál, and Lercher, 2004).

In *Saccharomyces cerevisiae*, consecutive gene pairs in the genome show a higher level of co-expression than widely separated genes (Cohen et al. 2000; Kruglyak and Tang 2000). These co-expressed genes cannot be operons, because the two genes often occur on opposite strands of DNA, making polycistronic transcription impossible (Cohen et al., 2000). The same co-expression of neighboring genes exists in *C. elegans*, which can largely be accounted for by operons, but which is still present in gene pairs that are not part of the same operon (Lercher, Blumenthal, and Hurst, 2003*a*). Higher-order levels of gene organization have also been discovered. For example, muscle-specific genes in *C. elegans* occur in blocks up to five genes in length (Roy, et al. 2002). In *Drosophila melanogaster*, even larger structures of gene organization are present, with 20% of genes organized into clusters with similar expression patterns ranging in size from 10 to 30 genes and up to 200 kb in length (Spellman and Rubin, 2002). In the mouse genome, both housekeeping and immunogenic genes have been found in clusters (Williams and Hurst, 2002). Clusters of housekeeping genes are also present in the human genome (Lercher, Urrutia, and Hurst 2002), in addition to clusters of highly expressed genes (Caron et al. 2001; Versteeg et al. 2003) and muscle-specific genes (Bortoluzzi et al. 1998). Thus, there is abundant evidence from a range of organisms that gene order in eukaryotic genomes is not random. But the reasons for this non-random arrangement are still unclear.

The co-expression of closely spaced genes might be attributable to chromatin structure (Hurst, Pál, and Lercher 2004). For example, it is known that when chromatin is opened to facilitate gene transcription, the open region can extend to neighboring genes (Stalder et al. 1980; Hebbes et al. 1994). Thus, the transcription of one gene could influence the transcription of neighboring genes, even if such a relationship is unintended (Spellman and Rubin 2002). Could natural selection be tolerating the co-expression of neighboring genes rather than actively promoting it? Two alternative hypotheses can explain the co-expression of neighboring genes. On the one hand, neutralist hypothesis might propose that the two genes are functionally unrelated but that *cis*-acting regulatory elements cause the transcription of one gene to influence the transcription of its neighbor. A selectionist hypothesis, on the other hand, might propose that co-regulation of these genes is required and that a chance rearrangement in the past brought them together (and thus facilitated their co-expression), which proved advantageous

[1] Present address: Human Cancer Genetics Program, The Ohio State University, Columbus, OH.

[2] Present address: University College Dublin, Dublin 4, Ireland.

[3] Present address: Center for Genomics and Bioinformatics, Karolinska Institutet Campus, Berzelius väg 35, SE-171 77 Stockholm, Sweden.

Key words: genome organization, human genome, mouse genome, natural selection.

E-mail: gacsinger@gmail.com

Mol. Biol. Evol. 22(3):767–775. 2005
doi:10.1093/molbev/msi062
Advance Access publication December 1, 2004

Molecular Biology and Evolution vol. 22 no. 3 © Society for Molecular Biology and Evolution 2004; all rights reserved.

enough for the new gene order to reach fixation in the population. One way of evaluating these alternative hypotheses is to use means other than expression data to define gene relationships. Lee and Sonnhammer (2003) have shown that genes involved in the same biochemical pathways tend to be clustered together in a variety of genomes, including human. Because the genes in these clusters were defined a priori as being co-regulated, the non-random grouping of these genes is hard to explain under the neutral model. Another way of distinguishing selection from neutrality in the co-expression of gene neighbors is to look for evidence of negative selection preserving the groups of genes over time. Indeed, this approach has shown that co-expressed gene pairs in *S. cerevisiae* are twice as likely to be preserved in *Candida albicans* as neighbors that are not co-expressed (Huynen, Snel, and Bork, 2001; Hurst, Williams, and Pál 2002), providing evidence that the gene pairings are an adaptation and not chance events. However, no studies have yet demonstrated that large blocks of co-expressed genes are preserved over the course of evolution.

Clusters of housekeeping genes are very prominent in both the mouse (Williams and Hurst 2002) and human genomes (Lercher, Urrutia, and Hurst 2002), but the orthology of these clusters has not been shown, nor have any studies measured the degree of preservation of these clusters relative to the rest of the genome. Here, we use microarray expression data from the Gene Expression Atlas (Su et al. 2002) to identify clusters of co-expressed and broadly expressed genes in the human and mouse genomes, confirming previous results based on expressed sequence tag (EST) and serial analysis of gene expression (SAGE) expression data (Lercher, Urrutia, and Hurst 2002). We then investigate whether human gene expression clusters remain chromosomal neighbors in mouse, and vice versa, and demonstrate that the clusters have been conserved to a greater degree than expected by chance. This indicates that natural selection is preserving the structure of these expression modules within each genome.

Methods

Expression Data

Gene expression data for mouse and human were taken from the Gene Expression Atlas (http://expression.gnf.org; Su et al. (2002)), which contains Affymetrix chip expression data (U74A for mouse, U95A for human) for many different tissues, 19 of which are common to both the mouse and human: adrenal gland, amygdala, cerebellum, cortex, dorsal root ganglia, heart, kidney, liver, lung, ovary, placenta, prostate, salivary gland, spleen, testis, thymus, thyroid, trachea, and uterus. Many of the expression experiments are replicated, and we took the mean expression for each tissue among the replicates. We eliminated genes that did not reach an Affymetrix Average Difference (AD) value of at least 200 in at least one tissue, and tissues for which the expression level was very low (AD values <100) were dropped to zero.

Mapping

UniGene clusters corresponding to Affymetrix tags were extracted from the Affymetrix probe consensus sequence file (http://www.affymetrix.com/support/technical/byproduct.affx?cat=arrays). In some cases, two or more Affymetrix tags were targeted against the same UniGene cluster, and only the tag with the highest average expression across all libraries from the human (or mouse) was retained.

UniGene clusters were mapped to the genome National Center for Biotechnology Information [NCBI] (build 31 and NCBIM build 30 for human and mouse, respectively) using the LocusLink and Ensembl databases. First, UniGene to LocusLink mapping was extracted from the UniGene release file Hs.data (human build U150) and Mm.data (mouse build U160). Second, LocusLink to Ensembl gene id mapping was extracted from the Ensembl database (release 14.31.1) using the EnsMart tool (http://www.ensembl.org/Multi/martview). LocusLink clusters mapping to multiple UniGene clusters or multiple Ensembl genes were discarded to ensure that the resulting mapping was unique and non-redundant. This procedure resulted in 4,451 human Affymetrix tags, and 4,522 mouse tags being mapped to the same number of unique locations on the human and mouse genomes. These sets of genes were used to infer the existence of clusters of genes with similar expression patterns. However, in the text we report the total number of genes within clusters, *including those for which we have no expression data*.

Removal of Duplicated Genes

Duplicated genes are expected to have similar expression patterns, and such genes are frequently located in physical proximity to each other and could give rise to a trivial clustering effect of co-expressed genes. We therefore removed all but one gene belonging to any gene family as determined by the TRIBE algorithm (Enright, Van Dongen, and Ouzounis 2002) and located within 10 Mb on the same chromosome. TRIBE families were extracted from the Ensembl database (release 14.31.1) using the EnsMart tool. After removal of duplicated genes, there were 4,114 human and 4,187 mouse Ensembl genes linked to Affymetrix tags.

Measures of Gene Expression Similarity

We used three measures of expression similarity in this study. First, for any two genes we measured the ''housekeepingness'' of the pair by multiplying the proportion of tissues in which gene *A* is expressed (AD value ≥ 200) by the proportion of tissues in which gene *B* is expressed. This measure has the advantage of being strongly skewed to the right, only assuming a high score if both genes are broadly expressed. This measurement was used to identify clusters of housekeeping genes. Second, we measured the height of expression for a pair of genes by taking the mean of their AD values across the 19 tissues listed in the *Expression Data* section, above. Previous studies have also used the median or maximum expression values (Caron et al. 2001; Versteeg et al. 2003), but these measures are all highly correlated with each other, so the choice between them is arbitrary (Lercher et al. 2003b). The height measure was used to search for clusters of highly expressed genes. Third, we used the Pearson correlation coefficient as a simple measure of co-expression

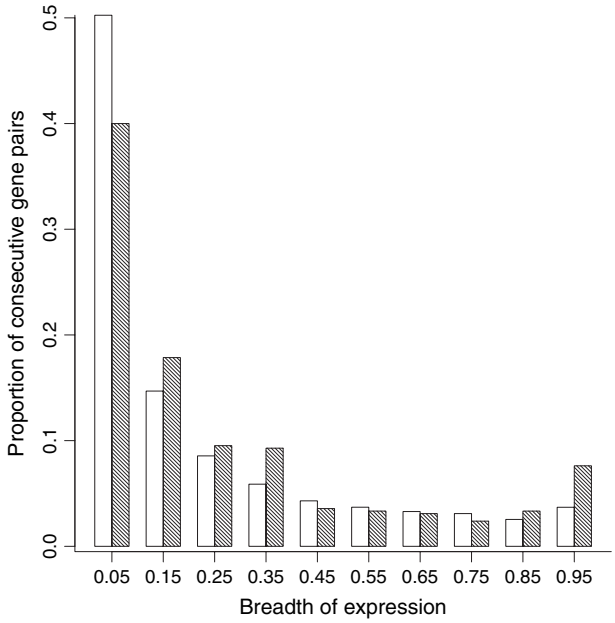


FIG. 1.—Distribution of expression breadth in random gene pairs (white bars) and consecutive gene pairs (shaded bars) in the human genome. There is an excess of broadly expressed (housekeeping) gene pairs in the consecutive pairs compared to the random pairs ($p < 0.0001$).

across the 19 tissues to search for clusters of co-expressed genes.

Sliding-Window Algorithm

To measure the clustering of gene expression patterns in the genome, we used a sliding-window analysis with a window size of 10 genes and a step size of one gene, ignoring genes for which no expression data were available. We limited the physical length of the windows by ignoring those that exceeded 0.5 Mb per gene in size. Within each window, the similarity in expression pattern (using each of the methods in the previous subsection) was measured between consecutive genes, and the scores for the consecutive pairs were then summed to form a score for the entire window. This score was compared to scores from 100,000 windows containing genes randomly sampled from the genome. Windows that had a similarity score exceeding that of 95% of the randomized windows were deemed significant. A second set of windows, with a more stringent significance cut-off of 99% were also annotated for later comparison to the more relaxed values. When the analyses were complete, overlapping significant windows were merged, forming blocks of ''clustered'' and ''unclustered'' regions in the genome.

Measuring Cluster Conservation

We identified mouse and human orthologous gene pairs using the EnsMart utility (version 14.1). Within each clustered (or unclustered) region of the genome, a chromosomal breakage ''opportunity'' (i.e., an intergenic region where an interchromosomal rearrangement could potentially occur during evolution) was counted between every two consecutive human genes whose mouse orthologs were known (any intervening human genes without known orthologs were ignored). If the mouse orthologs on either side of the breakage opportunity were on different chromosomes, a break event was counted. The number of breakage events relative to the number of opportunities was then compared between clustered and unclustered parts of the genome. We also compared the amount of intergenic space per chromosomal break within gene clusters to that in 1,000 randomized data sets, to control for the uneven spacing of genes within the genome.

Results

Pairs of Housekeeping Genes Are Maintained by Natural Selection

Previous reports based on SAGE and EST data have shown that housekeeping genes are clustered in the human (Lercher, Urrutia, and Hurst 2002) and mouse genomes (Williams and Hurst 2002). Unfortunately, the correlation between SAGE, EST, and microarray data tends to be very poor (Huminiecki, Lloyd, and Wolfe 2003), and it was therefore not clear that these clusters would be detectable using microarray expression data. For this reason, we began by using microarray data to confirm the existence of clusters of housekeeping genes in the human and mouse genomes. We initially looked for the smallest possible gene cluster: a simple pair of consecutive genes that have similar expression breadth across the 19 tissues. The number of physical gene pairs for which we have expression data for both genes is small, but it is sufficient for statistical analysis. When compared to 100,000 random gene pairs, it is clear that there is an excess of consecutive pairs with broad expression profiles in the human genome (fig. 1). The curve is shifted to the right in the consecutive gene pairs compared to the random pairs ($p = 1.75 \times 10^{-5}$ in a one-tailed Wilcoxon test), and there is a significant excess of gene pairs with an expression breadth exceeding 0.85 in the consecutive gene pairs (9.5% vs. 5.0% in the consecutive and random pairs, respectively; $p = 9.5 \times 10^{-5}$ in a one-tailed Fisher's exact test). Our examination of the mouse genome found similar results, with consecutive gene pairs having a higher similarity of expression breadth than the randomized gene pairs ($p = 0.00367$), and 6.9% of gene pairs exceeding the 95% level of expression breadth in the random gene pairs ($p = 0.029$). These findings indicate a significant level of organization of housekeeping genes in both the human and mouse genomes, at least at the level of gene pairs.

Although both genomes appear to contain linked pairs of housekeeping genes, it is possible that these pairings are an independent innovation within each lineage and are not orthologous. Moreover, it is necessary to distinguish chance pairs that exist in both the mouse and human genomes from those that have been maintained by purifying natural selection over time. To determine whether the pairs of housekeeping genes originated before the split of the mouse and human lineages and were conserved over the course of evolution, we downloaded orthology data from Ensembl to compare the proportion of conserved housekeeping gene pairs versus the proportion of conserved non-housekeeping pairs in each genome. We found that a very high proportion (29/30; 96.7%) of housekeeping pairs in human (defined as a breadth

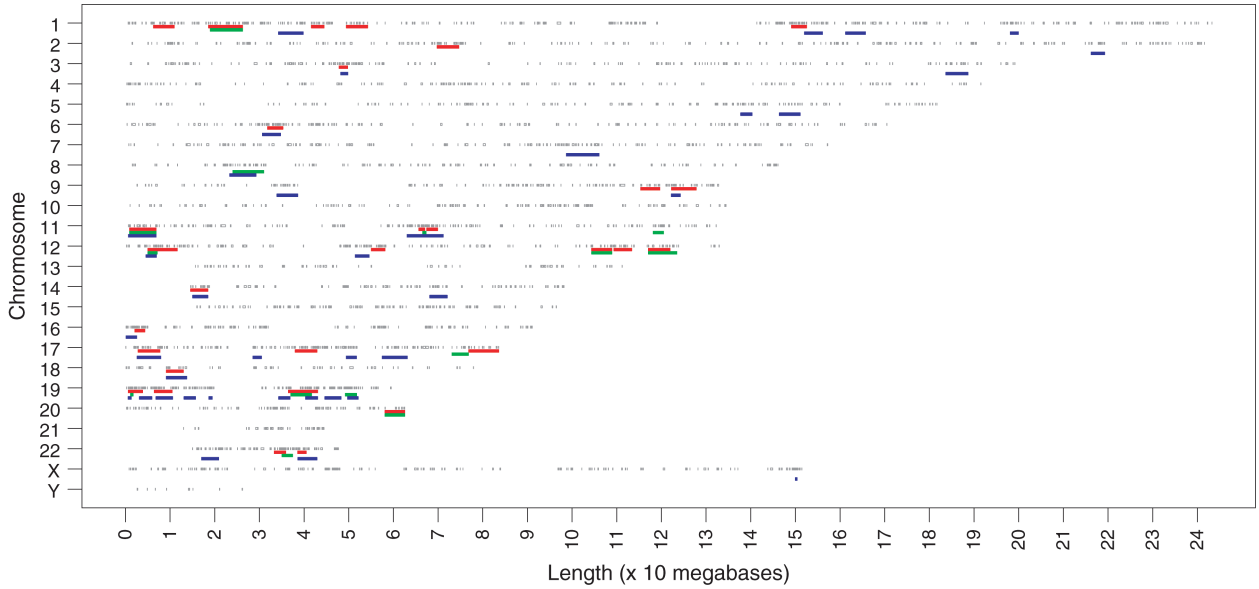


FIG. 2.—Maps of clusters of housekeeping (red), highly expressed (green), and co-expressed (blue) genes in the human genome. Tick marks indicate the locations of genes for which we have expression data. A larger version of this figure, as well as an equivalent diagram for mouse, is available in the Supplementary Materials online (figs. 1 and 2).

score of ≥0.85, corresponding to the upper ≈5% of random gene pairs) are also consecutive pairs in the mouse genome. This is significantly higher than the conservation of all gene pairs in the genome (84% (253/301); $p = 0.0175$ in a one-tailed exact unconditional test (Berger, 1996)). Similarly, 28/29 (96.6%) of housekeeping pairs in the mouse genome are within a one-gene distance of each other in the human genome, which is higher than that among the other genes for which expression and orthology data are known, where only 351/382 pairs (91.9%) are preserved (this difference is not significant, however: $p = 0.223$). These findings are suggestive that the pairing of these housekeeping genes is an adaptation that is maintained by purifying selection, and that they are not chance events.

Large Clusters of Housekeeping Genes Are also Maintained by Natural Selection

It is already known that clusters of housekeeping genes in the human genome extend beyond gene pairs (Lercher et al. 2002, 2003*b*), and we therefore wondered whether these larger structures were conserved over time. We performed a sliding-window analysis on both the human and mouse genomes to find clusters of housekeeping genes (see *Methods*). This procedure identified 30 housekeeping clusters in the human genome, containing between 33 and 203 genes, with a median of 86 genes (fig. 2). To assess whether these clusters are the product of natural selection or chance events, we examined the degree to which the arrangement of human genes in the clusters is conserved in the mouse genome. Of 1,567 opportunities for chromosomal breakage events (see *Methods*), 7 breaks were measured within the housekeeping clusters (0.44%). By comparison, 42 breakages of 4,303 opportunities were recorded outside the clusters (0.98%). This difference is small but statistically significant ($p = 0.031$ in a one-tailed Fisher's exact test). Our analysis of the mouse genome yielded similar results, where 30 clusters of housekeeping genes ranging in size from 39 to 127 genes were found, with a median of 62 genes (see figure 2 in the Supplementary Material online). As in the human genome analysis, we found a lower proportion of chromosomal breakpoints within the mouse housekeeping gene clusters than elsewhere in the genome (0.52% vs. 1.1%; $p=0.049$). We note that our test for cluster conservation only tests for interchromosomal breaks and does not test for conservation of local gene order within the clusters. It also does not require that the orthologs of clustered genes in one species should themselves form a cluster in the other species (the expression data are too sparse to allow for such stringency).

Previous studies have shown that clusters of co-expressed genes tend to be more tightly spaced than unclustered genes (Hurst, Williams, and Pál 2002), and that housekeeping genes in particular are short and closely spaced (Eisenberg and Levanon 2003). Indeed, we find that intergenic distances within our housekeeping clusters are much shorter than the average intergenic distance outside of clusters (the median intergenic distances within housekeeping clusters are 8,615 bp in human and 8,284 bp in mouse, less than half the values for the whole genomes). Because chromosomal breakage events are more likely between genes that are widely separated, the short distance between neighboring genes in housekeeping clusters could bias the outcome of our cluster conservation analysis. Therefore, we performed a randomization experiment in which gene locations were held constant but expression profiles for each gene were assigned randomly (without replacement). In that process, 1,000 randomized genomes were created, and the clustering analysis was performed on each one. It is clear that the number of genes involved in housekeeping clusters in the real genomes exceeds the numbers found in the randomized genomes (fig. 3a and 3b).

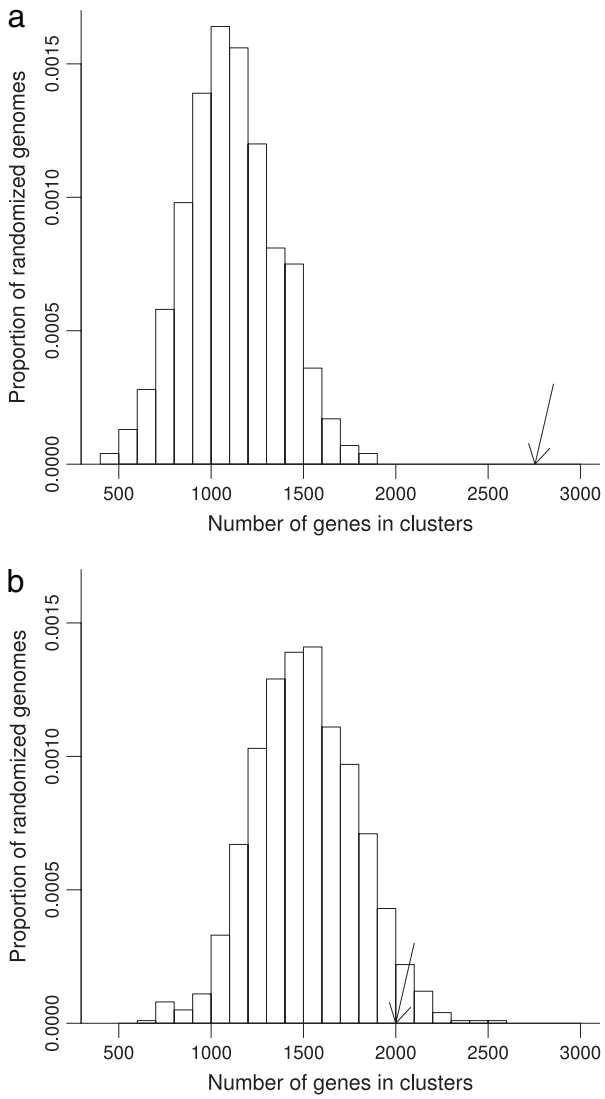


FIG. 3.—Distribution of the number of genes involved in housekeeping clusters in 1,000 randomized genomes for (a) human and (b) mouse. The numbers of genes found in housekeeping clusters in the real (non-randomized) genomes are indicated by arrows. Both are significantly higher than the means for the randomized datasets ($p < 0.04$).

Moreover, the real clusters are located in areas of the genome where interchromosomal rearrangements are relatively rare.

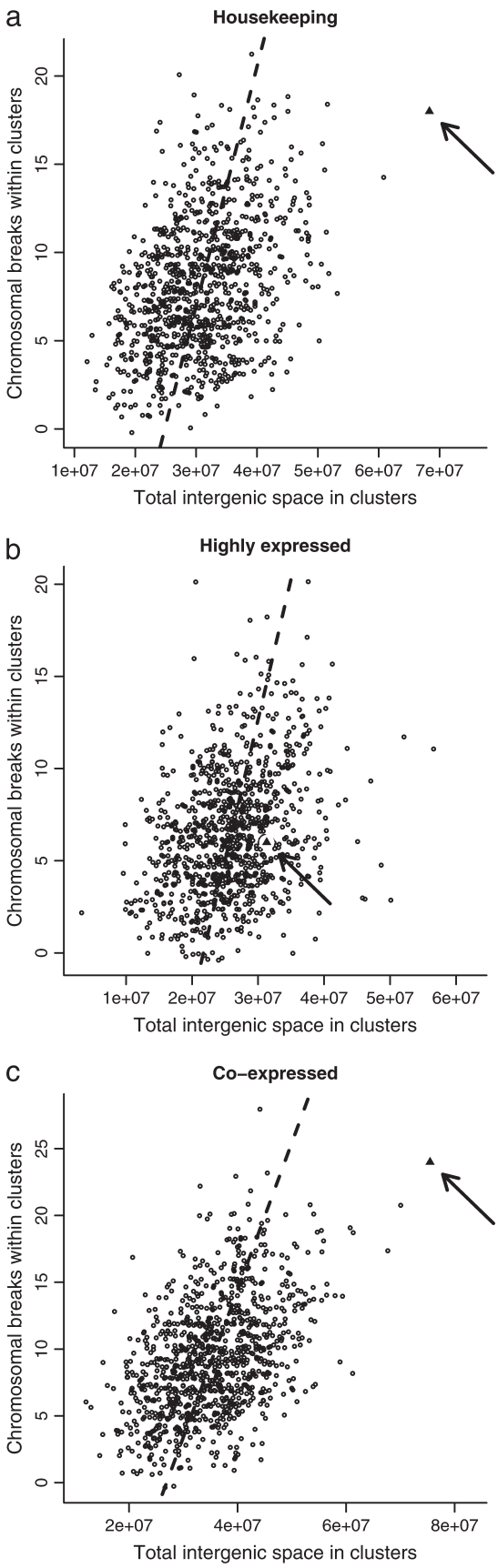
Considering the case of the human genome first, figure 4a shows a plot of the number of chromosomal breakages within housekeeping clusters versus the total intergenic spaces within those clusters within the 1,000 randomized genomes. The point for the real human genome (triangle in fig. 4a) clearly lies outside of the distribution of randomized points. With 18 chromosomal breakage events, the human housekeeping clusters should have a total of 38 mb of intergenic space based on a linear regression of the randomization data. However, they nearly double that size, at 68.3 mb, so these blocks are extraordinarily large considering the small number of chromosomal breakage events within them ($p = 0.001$). The equivalent analysis on the mouse genome reveals similar results (see figure 3a in the Supplementary Material online), where again the actual intergenic space within housekeeping clusters is greater than expected, given the number of chromosomal breakage events observed: with 11 chromosomal breaks, the clusters should contain 48.0 mb of intergenic space, but they actually contain 60.8 mb. This difference is not significant ($p = 0.078$), because the magnitude of deviation from expectation is less than observed in the human genome, although there is the suggestion of a tendency toward cluster conservation in the mouse genome.

Clusters of Highly Expressed Genes Are Not Maintained by Natural Selection

A number of studies have shown that highly expressed genes also form clusters within the genome (Caron et al. 2001; Versteeg et al. 2003). However, it is unclear whether this is simply an artifact of detecting clusters of housekeeping genes, because housekeeping genes tend to have above-average expression strength (Eisenberg and Levanon 2003; Lercher, Urrutia, and Hurst 2002). Indeed, in our data set there is a strong correlation between the height and breadth of expression for neighboring genes in the human genome (Spearman's $\rho = 0.71$), as well as in the mouse ($\rho = 0.69$). Nevertheless, we re-ran the sliding-window analysis, this time using expression height instead of breadth as the measure of expression distance between neighboring genes. We identified 14 clusters in the human genome, ranging in size from 29 to 176 genes. Of the genes in these clusters, 66% overlap with the housekeeping clusters identified previously (fig. 2). In mouse, only 12 clusters were found, with 24 to 131 genes in each cluster; 22% of the genes in these clusters are also found in mouse housekeeping clusters. We repeated the genome randomization procedure on both the mouse and human genomes, using expression height as our clustering attribute. In neither genome did the actual number of genes in clusters exceed that expected by chance. In human, the randomized genomes had 900 genes (standard deviation (σ) = 210) within highly expressed clusters on average, whereas the actual genome had 1,104 genes, a difference that is not statistically significant ($p = 0.17$ in a one-tailed Z-test). In the randomized mouse genomes, 1,210 genes ($\sigma = 226$) were placed within highly expressed clusters on average, whereas the actual genome had only 892 genes within highly expressed clusters ($p = 0.32$). Thus, we failed to find evidence of significant clustering of highly expressed genes at all. When we analyzed the degree of conservation of these clusters in both genomes using the methods described in the previous section, in neither case was the degree of conservation of the highly expressed gene clusters greater than what we expected by chance (see figure 4*b* for human and figure 3b in the Supplementary Material online for mouse).

Co-expressed Genes Are also Clustered in the Human and Mouse Genomes

Another measure of gene expression similarity used for detecting gene expression clusters is the simple Pearson correlation coefficient (Spellman and Rubin 2002; Stuart et al. 2003). An analysis of EST data has shown that linked genes in mouse tend to be co-expressed (Williams and Hurst 2002), although that study concluded that the effect was very weak. Our own analyses based on microarray data



identified large clusters of co-expressed genes in both mouse and human, and the number of genes involved in co-expression clusters greatly exceeds the numbers found in randomized genomes. In the human genome, our clustering algorithm placed 3,046 genes inside co-expression clusters, whereas the mean number in 1,000 randomized human genomes was only 1,281 ($\sigma = 291$). This difference is highly significant ($p = 5.78 \times 10^{-10}$ in a one-tailed Z-test). In mouse there were 2,455 genes in co-expression clusters, compared to an average of only 1,704 genes ($\sigma = 329$) in randomized genomes, which is also a significant excess ($p = 0.011$).

As with the housekeeping clusters, we compared the number of chromosomal breaks within the actual clusters identified in mouse and human to the number of chromosomal breaks within the clusters identified in randomized genomes. In both genomes, the number of breakage events is less than expected, given the amount of intergenic space within the clusters (see figure 4c for human and figure 3c in the Supplementary Material online for mouse). Thus, there is strong evidence that these co-expression clusters are being maintained by purifying selection.

Overlap Between Co-expression Clusters and Housekeeping Clusters

It is necessary to test whether co-expression and housekeeping clusters are independent of each other, or if there is significant overlap between the two. Visual inspection of figure 2 shows that, at least in the human genome, there does appear to be a significant overlap between the two measures. In fact, housekeeping and co-expression clusters share 37% of their genes in human and 20% in mouse. To see what effect these overlaps had on our previous analyses, we re-analyzed the housekeeping and co-expression clusters, this time removing genes that appeared in both clusters. In human, 1,656 genes are found within housekeeping clusters but not co-expression clusters, a number that is significantly higher than we observed when the same analysis was applied to randomized genomes (944, $\sigma = 229$; $p = 0.00094$). We also observed a higher degree of conservation of these clusters than we observed among the clusters from the randomized genomes (fig. 5a; $p = 0.012$). We observed the same trends in mouse, where again there was an excess of genes within housekeeping clusters (but outside co-expression clusters) compared to the numbers of genes found in the randomized mouse genomes (1,553 vs. 1,288, $\sigma = 259$). However, this result was not significant ($p = 0.15$). Moreover, these mouse clusters cannot be said to be conserved to a significantly higher degree than the randomized clusters, although again, the trend points in that direction ($p = 0.10$; see figure 4a of the Supplementary Material online).

←

FIG. 4.—Number of chromosomal break events versus the amount of intergenic space within expression clusters for 1,000 randomized human genomes. The triangle indicates the numbers for the nonrandomized human genome. Y-axis positions have been "jiggled" to reduce overlap between points. The true clusters are larger than expected ($p < 0.001$) based on the number of chromosomal breakage events observed for genes clustered by (a) expression breadth and (c) expression correlation, but clusters of highly expressed genes (b) lie well within the randomized data.

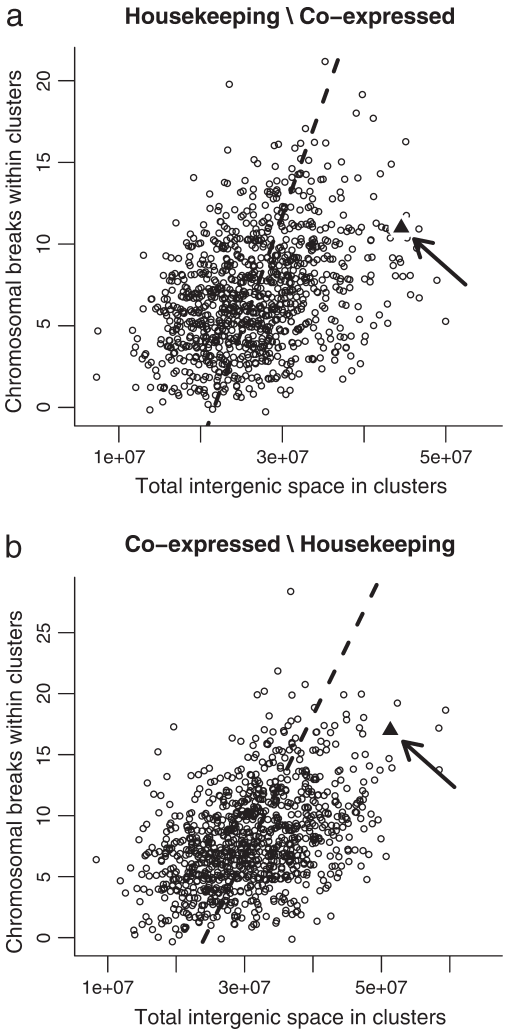


FIG. 5.—Number of chromosomal break events versus the amount of intergenic space within expression clusters for 1,000 randomized human genomes. Symbols and interpretation are the same as in figure 4, although in this case we have measured the degree of conservation of housekeeping gene clusters that do not overlap with co-expression clusters (a), and vice versa (b). In both cases, the true clusters are larger than expected (p = 0.012 and p = 0.0080, respectively).

We next did the reverse analysis, by re-analyzing the clusters of co-expressed genes after removing genes that also appeared within housekeeping clusters. Again, a significantly higher number of genes remained within co-expressed clusters than expected by chance (p=0.00082 for human and p=0.043 for mouse). The remaining clusters were found to be significantly conserved relative to randomizations in human (fig. 5b; p=0.0080), and mouse (see figure 4b of the Supplementary Material online; p=0.048).

Together, these two analyses show that housekeeping genes and co-expressed genes are independently clustered in the human and mouse genomes, and that these clusters are maintained by negative selection.

These Observations Are Robust to More Stringent Definitions of Gene Expression Clusters

It is worth noting that we used a very liberal significance cut-off when defining our gene expression clusters: a window of genes is considered significant if their similarity of expression (in terms of breadth, height, or correlation) exceeds the 95th percentile of 100,000 randomized gene windows (see the *Methods*). Because of the large number of windows scanned in the genome, this procedure will report many false positives. For this reason, we have repeated all of our analyses, only this time accepting gene windows whose measure of expression similarity exceeded the 99th percentile of randomized windows. As expected, the number of genes involved in expression clusters dropped significantly: only 966 genes are involved in housekeeping gene clusters in human, versus the 2,754 found using the less stringent cluster definition. Similarly, in mouse only 618 genes are found within housekeeping gene clusters using the strict cluster definitions, versus the 2,007 found when the relaxed definition was used. Nevertheless, these numbers still exceed the number of genes found in clusters in 1,000 randomized genomes, where the mean number of genes found within housekeeping gene clusters is 268 in human and 342 in mouse. Analogous results are found in the expression height clusters and co-expression clusters, where the results from the strict cluster definitions exactly mirror those from the loose definition, although the absolute number of genes is lower in all cases (table 1).

Table 1
Number of genes in clusters in real and randomized genomes, calculated using a stringent 99th-percentile definition of clusters

Organism	Cluster type	No. of genes in clusters	No. in bootstrapped genomes ± S.D.	Significance
Human	Houskeeping	966	268 ± 137	0.0010
	Highly expressed	74	197 ± 107	0.88
	Co-expressed	879	305 ± 144	0.0010
Mouse	Housekeeping	618	342 ± 146	0.046
	Highly expressed	268	234 ± 99	0.39
	Co-expressed	926	411 ± 156	0.0030

We also analyzed the degree of conservation of these stringently defined clusters and found that again, the patterns mirrored those of the liberally defined clusters: housekeeping gene clusters are maintained to a greater degree than clusters found within randomized genomes (p = 0.011 and p = 0.036 in human and mouse, respectively). This is also true of clusters of co-expressed genes (p = 0.050 and p = 0.0020 in human and mouse, respectively), and as expected, there is no evidence for conservation of clusters of highly expressed genes in either genome (p = 0.96 and p = 0.55 in human and mouse, respectively). Thus, the results are qualitatively identical regardless of how stringent we are in defining similarly expressed gene clusters: there is strong evidence for housekeeping gene and co-expressed gene clusters in both the mouse and human genomes, but not for highly expressed gene clusters.

Discussion

We have found clusters of housekeeping genes in the human genome by using microarray expression data, which confirms previous findings based on EST and SAGE data

(Lercher, Urrutia, and Hurst 2002). Interestingly, there are conspicuously few clusters on the X chromosomes in both mouse and human. We cannot rule out the possibility that this an artifactual result of the sparse data we have analyzed (the density of available data for chromosome X is lower than average in both organisms), but it is tempting to think that this is a real phenomenon, perhaps related to the decreased recombination on sex chromosomes which might result in a decreased opportunity for cluster formation. Although housekeeping gene clusters were previously known to exist in the mouse genome (Williams and Hurst 2002), the orthology of the clusters in the two genomes had not been demonstrated. We have shown that housekeeping clusters in the human genome tend not to be broken up in the mouse genome (and vice versa) relative to other groups of genes, lending support to the hypothesis that these clusters are advantageous and are therefore being preserved by purifying selection. It is notable that we found lower statistical support in all our analyses of the mouse genome than in the human genome. However, given the spotty nature of our data set (covering only about 20% of the genes in each genome), it is possible that this trend is a sampling artifact.

Our analyses suggest that reports of the clustering of highly expressed genes in the human genome (Caron et al. 2001; Versteeg et al. 2003) may, in fact, be indirectly detecting some of the clusters of housekeeping genes because there is a high degree of correlation between the two measures (see figure 2 and figure 2 of the Supplementary material online). This agrees with previous findings (Lercher, Urrutia, and Hurst 2002), but we have also demonstrated that, unlike clusters of housekeeping or co-expressed genes, clusters defined by expression height are not conserved to a greater degree than expected by chance, and are therefore probably not maintained by natural selection.

We have presented results indicating that co-expressed genes are clustered in both the mouse and human genomes, and that these clusters may be conserved by natural selection. Interestingly, previous reports have found that the clustering of co-expressed genes is a weak effect in the mouse (Williams and Hurst 2002) and human (Lercher, Urrutia, and Hurst 2002) genomes. Whether this disagreement is due to the expression data used (EST and SAGE data in other studies versus microarray data here), the genes analyzed (no study of this sort has sampled more than about 20% of known mouse and human genes), or the method used to define gene co-expression is unclear.

Our results show that purifying selection is preserving clusters of both housekeeping genes and co-expressed genes in the human genome, and that the same forces may be at work in the mouse genome. In most of our analyses, gene cluster conservation was observed to be weaker in mouse than in human. Although the trends in mouse were identical to those in human, they often did not reach statistical significance. Unfortunately, we cannot rule out the possibility that this is an artifact of the data we have analyzed. However, another possibility is that this is a real biological phenomenon. Given that the mouse genome has undergone a large amount of gene rearrangement (Mullins and Mullins 2004), perhaps gene clusters are being eroded over time in mouse, while at the same time they are being actively preserved in human. Although this scenario seems unlikely, it is consistent with our results, and further study will be needed to rule it out.

What is the advantage of gene clustering in the first place? Presumably, the close proximity of the genes is an adaptation that facilitates the co-regulation of their transcription. Eisenberg and Levanon (2003) reported that the Gene Ontology annotations (Ashburner et al. 2000) for human housekeeping genes show a high proportion of metabolism-related and RNA-interacting proteins (such as ribosomal proteins). These types of genes play a fundamental role in the operation of every eukaryotic cell, and thus, if there is any benefit to arranging co-expressed genes together in the genome, housekeeping genes will likely be subject to the strongest selection coefficients to form such clusters. Moreover, housekeeping genes tend not only to be broadly expressed but also highly expressed, which is a pattern that probably requires little regulation in comparison to genes with very specific expression patterns. Perhaps housekeeping genes are more amenable to being controlled by broadly acting *cis*-regulatory elements than other genes, or perhaps they are subject to repression of transcription through chromatin modification in particular circumstances.

In conclusion, we have shown that there appears to be a selective benefit to the clustering of co-expressed and broadly expressed genes in the human and mouse genomes. We believe this is the strongest evidence to date that the non-random arrangement of genes in mammalian genomes is the product of natural selection.

Acknowledgments

This study was supported by a Science Foundation Ireland grant to K.H.W. We thank the two referees of this paper for their constructive comments.

Literature Cited

Ashburner, M., C. A. Ball, J. A. Blake et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

Berger, R. L. 1996. More powerful tests from confidence interval p values. Am. Statisti. **50**:314–318.

Blumenthal, T. 1998. Gene clusters and polycistronic transcription in eukaryotes. Bioessays **20**:480–487.

Blumenthal, T., D. Evans, C. D. Link, et al. 2002. A global analysis of Caenorhabditis elegans operons. Nature **417**:851–854.

Bortoluzzi, S., L. Rampoldi, B. Simionati, et al. 1998. A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res. **8**:817–825.

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

Cohen, B. A., R. D. Mitra, J. D. Hughes, and G. M. Church. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat. Genet. **26**:183–186.

Eisenberg, E., and E. Y. Levanon. 2003. Human housekeeping genes are compact. Trends Genet. **19**:362–365.

Enright, A. J., S. Van Dongen, and C. A. Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30**:1575–1584.

Hebbes, T. R., A. L. Clayton, A. W. Thorne, and C. Crane-Robinson. 1994. Core histone hyper-acetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. EMBO J. **13**:1823–1830.

Huminiecki, L., A. T. Lloyd, and K. H. Wolfe. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. BMC. Genomics **4**:31.

Hurst, L. D., C. Pál, and M. J. Lercher. 2004. The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5**:299–310.

Hurst, L. D., E. J. Williams, and C. Pál. 2002. Natural selection promotes the conservation of linkage of co-expressed genes. Trends Genet. **18**:604–606.

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

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

Lee, J. M., and E. L. Sonnhammer. 2003. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. **13**:875–882.

Lercher, M. J., T. Blumenthal, and L. D. Hurst. 2003*a*. Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes. Genome Res. **13**:238–243.

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

Lercher, M. J., A. O. Urrutia, A. Pavlícek, and L. D. Hurst. 2003*b*. A unification of mosaic structures in the human genome. Hum. Mol. Genet. **12**:2411–2415.

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

Niehrs, C., and N. Pollet. 1999. Synexpression groups in eukaryotes. Nature **402**:483–487.

Roy, P. J., J. M. Stuart, J. Lund, and S. K. Kim. 2002. Chromosomal clustering of muscle-expressed genes in Caenorhabditis elegans. Nature **418**:975–979.

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

Stalder, J., M. Groudine, J. B. Dodgson, J. D. Engel, and H. Weintraub. 1980. Hb switching in chickens. Cell **19**:973–980.

Stuart, J. M., E. Segal, D. Koller, and S. K. Kim. 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science **302**:249–255.

Su, A. I., M. P. Cooke, K. A. Ching, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. Proc. Natl. Acad. Sci. USA **99**:4465–4470.

Versteeg, R., B. D. Van Schaik, M. F. Van Batenburg, M. Roos, R. Monajemi, H. Caron, H. J. Bussemaker, and A. H. Van Kampen. 2003. The Human Transcriptome Map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res. **13**:1998–2004.

Williams, E. J., and L. D. Hurst. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. J. Mol. Evol. **54**:511–518.

Zorio, D. A., N. N. Cheng, T. Blumenthal, and J. Spieth. 1994. Operons as a common form of chromosomal organization in C. elegans. Nature **372**:270–272.

William Martin, Associate Editor

Accepted November 25, 2004