J Mol Evol (2002) 54:511–518
DOI: 10.1007/s00239-001-0043-8

JOURNAL OF MOLECULAR EVOLUTION

Clustering of Tissue-Specific Genes Underlies Much of the Similarity in Rates of Protein Evolution of Linked Genes

Elizabeth J.B. Williams, Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

Received: 27 August 2001 / Accepted: 1 October 2001

Abstract. Are genes nonrandomly distributed around the genome and might this explain why it was found that, in the mouse genome, proteins of linked genes evolve at similar rates? Anecdotal evidence suggests that the similarity of expression of linked genes might, in part, explain the similarity in their rates of evolution. Immune system genes, for example, are known to evolve at a high rate and sometimes cluster in the genome. Here we develop methods for statistical tests of similarity of expression of linked genes and report that there is a significant tendency for genes of similar expression breadth to be linked. Significantly, when we exclude tissue specific genes from our sample, the similarity in rates of protein evolution of linked genes is greatly diminished, if not abolished. This diminution is not a sampling artifact. In contrast, while half of the immune genes in our sample reside in 1 of 10 immune clusters in the mouse genome, this clustering appears not to affect the extent of local similarity in rates of evolution. The distribution of placentally expressed genes, in contrast, does have an effect.

Key words: Expression patterns — Linkage — Rate of evolution — MHC

Introduction

Are genomes strings of genes in no particular order or might it be the case that selection favors certain genes to be clustered, possibly to ensure coregulation? While operon structures are well described in bacteria, the linkage of coexpressed genes in eukaryotes is typically considered the exception rather than the rule. However, this view might be changing. In the human genome highly expressed genes appear to be clustered (Caron et al. 2001). Similarly, recent systematic evidence indicates that skeletal muscle genes (Bortoluzzi et al. 1998), extraembryonically expressed genes (Ko et al. 1998), olfactory genes (Lander et al. 2001), and tRNA genes (Lander et al. 2001) tend to show clustering (although only the analysis of extraembryonic genes controls for tandem duplication). Likewise, genes in the MHC cluster tend to be involved in immune functions, and in some cases the most tightly linked (e.g., *Tap* and *LMP*) are involved in coupled processes (Hughes and Yeager 1997).

Here we compile data on expression profiles of a few hundred mouse genes, of known genomic location, to ask whether similarly expressed genes tend to be linked more often than expected by chance. To achieve this we develop measures of similarity of expression. In particular, we examine (1) the breadth of expression, meaning the number of tissues in which a gene is expressed, and (2) the degree of coexpression, meaning the correspondence between genes in the degree to which they are expressed in the same tissues. These two are logically distinct, as two tissue specific genes, for example, will show similar expression breadth but may be expressed in different tissues (i.e., no coexpression). Additionally, we examine a specific coexpression hypothesis. Given that genes in the MHC tend to be immune related, we ask whether

Correspondence to: Laurence D. Hurst; *email:* l.d.hurst@bath.ac.uk

immune system genes tend to be clustered more often than expected by chance and whether the MHC might be the exception or the rule.

Expression, Linkage, and Rates of Evolution

The motivation behind these tests is not simply to allow a better statistical appreciation of the degree of ordering in the mouse genome. We also wish to understand whether such patterns might underpin the recently described similarity in the rate of evolution of the proteins of linked genes (Williams and Hurst 2000). For this to be so there needs to be a relationship among expression pattern, linkage, and rates of protein evolution.

Evidence that expression pattern (broadly defined) might be related to the rate of protein evolution comes from a variety of sources. Importantly, proteins of genes expressed in a tissue-specific manner evolve on average twice as fast as those that are ubiquitously expressed (Duret and Mouchiroud 2000). Further, the proteins of certain tissues tend to evolve faster than others. Most notably, immune system genes evolve about twice as fast as nonimmune genes (Hurst and Smith 1999). It is for this reason that we wish to examine the spatial genomic distribution of immune system genes in particular.

Methods

Data Set Compilation

We compiled a data set of mouse and rat orthologues from scrutiny of entries in HOVERGEN (Duret et al. 1994). Genes were accepted as orthologues if, and only if, the mouse and rat sequences had no other nonrodent sequence separating their branches and at least one nonrodent sequence appeared as a sister group. This resulted in a data set of in excess of 500 gene pairs.

Each of the mouse genes was then inspected at LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), using its accession number, to establish mouse chromosomal location. These chromosomal locations are the same as those described at Mouse Genome Informatics (http://www.informatics.jax.org/). Only autosomal genes with a location specified to the centimorgan (cM) were used, because X-linked genes have unusually low rates of evolution (Smith and Hurst 1999). Pairwise Blast under the default settings was used to eliminate tandem duplicates from the data set. Any reported similarity between linked genes led to the elimination of one of the two. This resulted in a data set of 475 autosomal genes. Of these, 289 had at least one neighbor within 1 cM.

Molecular Evolutionary Analysis

The coding sequence was extracted automatically using the annotations in the GenBank entry. DNA alignments were carried out by PILEUP (Wisconsin Package, GCG) using the default settings. The alignments were checked by eye and modified if the alignment was obviously wrong (e.g., translation of aligned sequences gave a nonfunctional protein). Substitution rates were estimated using the method described by Li (1993; Pamilo and Bianchi 1993), applying Kimura's two-parameter method to correct for multiple hits, and by the maximum likelihood method of Goldman and Yang (1994). For each orthologous gene (mouse–rat) we therefore obtained two estimates for the rate, per site, for both nonsynonymous (K_a) and synonymous (K_s) substitutions. We also calculated the rate of protein evolution, controlling for the underlying mutation rate (K_a/K_s). However, we have found that none of the results that we present below are greatly affected by the choice of method. Therefore, for ease of comparison we report only the results using Li's protocol, except where of unusual interest (precise figures for results using the maximum likelihood protocol available on request).

Expression Data

Expression data were assembled from numerous resources. First, all genes were inspected at Unigene (http://www.ncbi.nlm.nih.gov/UniGene/) and the tissues of confirmed expression were noted. These data are based on EST matches of genes and will give only a positive result; negative results are not reported. Additionally, the expression data given at MGI (http://www.informatics.jax.org) was employed. Finally, the original source papers were consulted. If there is disagreement between or within the resources whether a gene is or is not expressed in a certain tissue, we always count the gene as being expressed, under the supposition that a false positive is considerably less likely than a false negative.

From the source papers we could classify some genes as definitely not being expressed in certain tissues (at least at certain times and in certain strains). When a tissue was actively investigated for expression but none was found, we refer to this as the narrow definition of nonexpression. Using this methodology we can, for each gene, score the expression in any given tissue as present, not present (from narrow definition), or "no hit" (not a clear positive or negative due to lack of firm data).

Twenty-two tissues were considered. For each gene, we can obtain a score for the total number of tissues in which expression has been reported. This we define as the breadth of expression. While in principle this value might run from 0 to 22 (no expression to ubiquitous expression), we eliminated all those scoring 0, regarding it as evidence that the expression of the gene has yet to be adequately investigated.

Index of Coexpression (ICE)

Not only can we calculate the breadth of expression, but also we can calculate the degree of coexpression for any given pair of genes. This index of coexpression was calculated as follows. If in a given tissue both genes were expressed, or both were not expressed, then the gene pair scores one for that tissue. Expression of one gene and not the other gives a score of −1. This procedure was followed for each of the 22 tissues and a total score was calculated. This total was then divided by the total number of informative tissues to provide an index of coexpression (ICE) that can run from −1 to +1. An ICE value of +1 means perfect coexpression; both genes were expressed in the same tissues and only those tissues. A negative ICE implies antagonistic expression, i.e., where one gene was expressed, the other was not. An ICE value of 0 means coexpression half the time and antagonistic expression the other half. The definition of an "informative tissue" and of "no expression" depends on the precise model that we use. These we now outline.

Models for ICE

We employed three models that differed in their interpretation of the "no hit" category of expression and how this relates to nonexpression. As the data are derived from matches to EST data, it is not the case that no hit simply means no information; it might indicate absence of expression.

Table 1. Summary of the p and r^2 values obtained using the randomization protocols (a) devised by Lercher et al. (2001)[a] and (b) used by Williams and Hurst (2000)[b]

	No. genes	No. comparisons	K_a		K_a/K_s	
			p	r^2	p	r^2
(a)						
Whole data set	289	223	0.0029	7.2%	0.011	10.6%
Without immune genes	243	181	0.053	7.9%	0.12	8.1%
Tissue-specific genes	134	76	0.034	14.0%	0.031	11.9%
Tissue-specific without immune genes	80	51	0.087	13.3%	0.19	10.2%
Without tissue-specific genes	155	87	0.54	0.0%	0.081	5.3%
Tissue-specific without placentally expressed genes	127	67	0.125	9.5%	0.073	9.3%
(b)						
Whole data set	289	196	0.0001	5.8%	0.0001	5.6%
Without immune genes	243	147	0.02	4.1%	0.008	6.2%
Tissue-specific genes	134	61	0.0061	9.6%	0.013	8.1%
Tissue-specific without immune genes	80	38	0.0094	20.5%	0.0083	16.4%
Without tissue-specific genes	155	74	0.34	0.2%	0.1788	2.1%
Tissue-specific without placentally expressed genes	127	55	0.078	5.0%	0.0501	5.8%

[a] These were obtained by comparing each individual gene's K_a and K_a/K_s values with the average of its neighbors. The p value was obtained from randomizations, and the r^2 value from linear correlation.

[b] These were obtained by pairing linked genes using no gene more than twice in total. The p value was obtained from randomization of the mean difference in K values between the pairs of linked genes. The r^2 value was obtained from linear correlation of the K values of the linked genes.

Model 1: No Hit = No Information. At one extreme we can suppose, conservatively, that "no hit" is synonymous with an absence of information. This is reflected in the calculation of the index of coexpression that we calculate for all pairs of linked genes. In this model, informative tissues are those in which expression is either present or confirmed absent for both genes in the pair. If either gene has a no hit, this is treated as an absence of data so is not counted as an informative tissue. When calculating the mean index in any given set of gene pairs, we calculated a mean weighted by the total number of informative tissues.

This method has the problem that it is biased to reporting high ICE values, as most of the information available confirms the presence of expression. An extreme example is that if there were no confirmed lack of expression, all genes would score either 0 for no matches or 1 for at least one confirmed match.

Model 2: No Hit = No Expression. At the other extreme we can suppose that "no hit" means no expression, in which case the number of informative tissues is always 22. This model tends to report high ICE values when the number of no hits is high. Tissues ignored in model 1 because both genes scored no hit, will now return a +1 value to the score. If the sampling of expression data is extensive and EST matches are well reported, then this should, in principle, provide the most reliable information. However, if the sampling is sparse (as must be the case to some extent if some genes have failed to be detected at all or some tissues are not used extensively in EST studies), then this overestimates the degree of coexpression.

Model 3: A Hybrid Model. In our hybrid model we assume that a "no hit" counts as no expression, but only if in the tissue concerned the other gene in the pair has a confirmed expression pattern. This hybrid model attempts to minimize the effect of poor data on the ICE values in model 2. That is, there may be several data points for any given gene pair that score +1 simply because there are many no hit results. If this is due to poor data, rather than a true reflection of expression, this is a problem. In this model the number of informative tissues is 22 minus the number of tissues where both genes have a no hit.

Statistical Analysis

A randomization protocol was used to analyze how similar the expression profiles of linked genes were. To analyze the extent to which linked genes had similar breadths of expression, for each linked pair we calculated the difference in the total number of tissues in which each gene was expressed. This value was calculated for all the linked gene pairs in the data set, and the mean calculated. This mean was compared with means calculated from 10,000 randomizations of the data set. In the randomizations the expression profile of the genes were conserved and the gene position in the genome was randomized. If a gene was used more than once in the original data, it was used more than once in the randomized data set.

We performed a similar procedure for the analysis of the index of coexpression by each of the three models. We calculated a mean (or weighted mean) index of coexpression for the real data. We then randomized the gene positions and calculated a mean index of coexpression for 10,000 randomized sets. These methods allow us to ask how often we would expect by chance the degree of similarity of expression profiles of linked genes which we obtained from the real data set.

For calculating the similarity in rates of evolution we used the method developed by Lercher et. al. (2001) and that employed by Williams and Hurst (2000). The former differs from the previous randomization protocols as it calculates, for each gene, the mean difference between the gene's value (K_a or K_a/K_s) and the mean value of all its neighboring genes within 1 cM. The mean difference calculated from the real data set is then compared to a set of 100,000 random mean values calculated in the same way from randomized data sets. For each test we report (Table 1a) the p value and the r^2 value. The latter is calculated by correlating each individual gene's K value with the mean of its neighbors. In Table 1b we also report the results using the method used by Williams and Hurst (2000) as a comparison. In this method a given gene is compared to its nearest two neighbors (or one neighbor if only one other is within 1 cM). Often, however, the choice of nearest neighbor is arbitrary, as recombination maps place many genes at the same position. The results obtained are sensitive to methodology (Lercher et al. 2001). Given the slightly arbitrary nature of the method

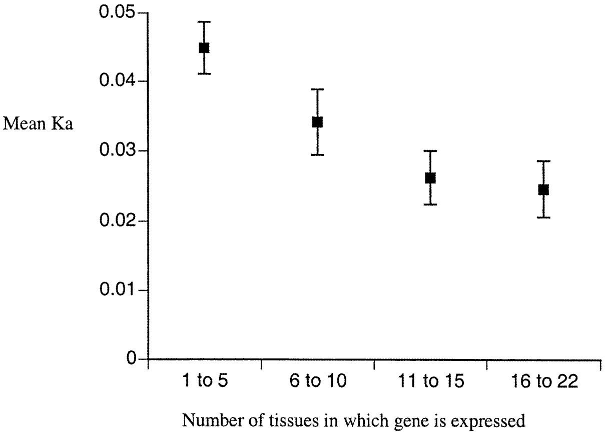

Fig. 1. The relationship between the number of tissues in which a gene is expressed and its rate of nonsynonymous evolution: 1–5 tissues, $N = 133$; 6–10 tissues, $N = 64$; 11–15 tissues, $N = 56$; 16–22 tissues, $N = 32$.

used by Williams and Hurst (2000), we report in the text the results obtained by the method of Lercher et al. (2001) while flagging up results that appear to be method sensitive.

Results

Proteins of Highly Expressed Genes Are Slow-Evolving

We can confirm that within our data set the breadth of gene expression (E) is negatively correlated with the rate of protein evolution ($K_a = 0.046 - 0.0014E$, $r^2 = 3.9\%$, $p < 0.001$) (Fig. 1). Note, however, that this is a relatively weak effect. As with previous analysis, we find no evidence that K_s and expression level covary ($K_s = 0.00018 - 0.0003E$, $r^2 = 0.1\%$, $p = 0.564$).

Proteins of Linked Genes Evolve at a Similar Rate

We can confirm in this data set that the difference in K_a between linked genes is much lower than expected by chance ($p = 0.0029$, $r^2 = 7.2\%$) (Table 1a). Parenthetically, as regards local similarity in K_s, we previously reported (Williams and Hurst 2000) weak significance ($p = 0.01$). In the present data set this effect has decreased marginally ($p = 0.039$, $r^2 = 1.5\%$).

Linked Genes Have a Weak Tendency to Have Similar Expression Patterns

To ask whether linked genes might show similar expression patterns we analyzed the local similarity of expression profiles using two measures.

Expression Breadth. To investigate whether linked genes had similar expression breadths, we calculated the mean difference in breadth of expression (calculated as the total number of tissues in which each gene is expressed) of linked genes and compared this with the mean from 10,000 randomized simulants. We find that only 4% of randomized data sets show a higher level of local similarity in breadth of expression. A priori we would expect that 50% of random data sets would show a higher level of local similarity in expression breadth, therefore this result shows that there is a significant tendency for linked genes to have similar expression breadths.

Degree of Coexpression. Coexpression of linked genes was investigated using the three ICE (index of coexpression) models (explained in methods) for the interpretation of the expression data. Again, we compared the mean (or weighted mean) ICE with the distribution of ICE values obtained through randomization. In each we find at most a weak tendency for linked genes to be more similarly expressed than expected by chance: Model 1 (“no hit” = no information), $p = 0.095$; Model 2 (no hit = no expression), $p = 0.183$; and Model 3 (hybrid model), $p = 0.093$.

Clustering of Immune System Genes Is Very Common. The above results suggest that clusters of genes expressed in the same tissues are the exception rather than the rule. But is this also true if we look more specifically at immune system genes? For these we have a priori expectations that they might be clustered given the presence of the MHC cluster. However, it is hard to provide a definitive definition of what is and what is not an “immune system gene.” We chose to apply a method that takes account of as much information as possible. We therefore used all available functional information and expression data and defined a gene as being of the immune system if (a) the knockout had an effect on the immune response or (b) it had expression specific to immune cells (e.g., B cells and T cells). Additionally, Mouse Genome Informatics defines certain genes as belonging to the immune system. We included any gene that MGI considered as belonging to the immune system. No doubt one might query whether our definition is too conservative or too liberal, but in the absence of alternative objective criteria and definitions, we consider this to be a reasonable approach and not obviously prone to bias.

In our data set we find strong evidence for clustering of unrelated immune system genes. There are 46 immune system genes, 24 of which have at least one other immune gene within 1 cM. These exist in 10 clusters, 2 of which are relatively large (Table 2). We could define 13 pairs of linked immune system genes. In 10,000 randomized data sets, on the average there are only 3.75 linked immune pairs (and a maximum of 11). The frequency of

Table 2. The 10 clusters of immune system genes and their chromosomal locations for genes within our sample[a]

Name of gene	Chromosome	cM position	K_a	K_s	K_a/K_s
Interleukin 1 receptor, type I	1	19.5	0.08	0.273	0.293
Interleukin 1 receptor, type II	1	19.5	0.054	0.162	0.333
CD28 antigen	1	30.1	0.055	0.279	0.197
CD152 antigen CTLA	1	30.1	0.046	0.137	0.336
Decay accelerating factor 1	1	67.6	0.185	0.234	0.791
Polymeric immunoglobulin receptor	1	68.5	0.075	0.176	0.426
Cathepsin E	1	69.1	0.036	0.161	0.224
Interleukin 10	1	69.9	0.077	0.173	0.446
Selectin, platelet	1	86.6	0.054	0.228	0.237
CD3 antigen, ζ polypeptide	1	87.2	0.034	0.148	0.230
CD1d1 antigen	3	48.0	0.087	0.21	0.414
CD53 antigen	3	48.5	0.038	0.173	0.220
Small inducible cytokine B subfamily (Cys–X–Cys), mbr 10	5	53.0	0.129	0.329	0.392
Small inducible cytokine B subfamily, mbr 5	5	53.0	0.115	0.241	0.477
CD9 antigen	6	57.0	0.032	0.166	0.193
Tumor necrosis factor receptor superfamily, mbr 1a	6	57.1	0.092	0.184	0.50
Chemokine (C–C) receptor 1, -like 2	9	72.0	0.042	0.143	0.293
Chemokine (C–C) receptor 2	9	72.0	0.031	0.136	0.228
Small inducible cytokine A2	11	46.5	0.098	0.099	0.989
Small inducible cytokine A11	11	47.0	0.025	0.062	0.403
Small inducible cytokine A5	11	47.0	0.023	0.115	0.200
Small inducible cytokine A3	11	47.6	0.064	0.145	0.441
Histocompatibility 2, class II, locus DMa	17	18.56	0.069	0.229	0.301
Tumor necrosis factor	17	19.06	0.035	0.157	0.223

[a] A cluster is defined as the presence of one or more immune system genes within 1 cM of another immune gene. Also listed are the rates of nonsynonymous (K_a) and synonymous evolution (K_s). For the data set as a whole the mean K_a is 0.04 ± 0.04 and the mean K_s is 0.174 ± 0.05. The mean K_a/K_s for these genes is 0.21 ± 0.21, but for these linked immune system genes it is 0.39 ± 0.12.

linked immune system genes is therefore significantly higher than expected by chance ($p < 0.0001$).

Clustering and Rates of Protein Evolution

The above set of results presents highly contrasting pictures: broad-scale analyses report only weak effects, at most, for the effects of linkage on similarity of expression. In contrast, within the same data sets there is a strong pattern of clustering of immune system genes, an effect that is diluted in the broad-scale pattern. A priori given the weakness of the broad-scale patterns, it seems unlikely that a broad-scale analysis of the extent to which expression similarity covaries with the similarity of rates of evolution of linked genes will provide an informative result. However, this seems not to be the case.

Clustering of Tissue-Specific Genes Underlies the Local Similarity in Rates of Evolution. A gene can be defined as being tissue specific if it is expressed in fewer than six tissues. This, naturally, is an arbitrary divide. However, genes expressed in only one tissue are too few to provide meaningful analysis. We can then partition our sample into tissue-specific ($N = 134$) and nonspecific ($N = 155$) genes. Within the nonspecific group we find that the local similarity is removed completely (K_a: $p = 0.54$, $r^2 = 0.0\%$) (Table 1a, Fig. 2). In contrast, when we look at tissue-specific genes and test for local similarity in rates of evolution, we find that there is a correlation stronger than before, even though the p value does not indicate that it is highly statistically significant (K_a: $p = 0.034$, $r^2 = 14.0\%$) (Table 1a, Fig. 2). Using the method of Williams and Hurst (2000), the relevant p value resolves to 0.0061 ($r^2 = 9.6\%$), versus $p = 0.0001$ ($r^2 = 5.8\%$) for the data set as a whole. This suggests that the local similarity in the rate of protein evolution is due largely to the distribution and rate of evolution of tissue-specific genes.

Could the apparent absence of local similarity in the non-tissue-specific set of genes be an artifact of dividing the data set up, thereby reducing the sample size? To examine this, we randomly divided the entire data set into two subsamples, one the same size as the tissue-specific group ($N = 134$) and the other containing the remainder ($N = 155$). We repeated this 100 times. We then calculated the extent of local similarity in each random subsample using the method of Lercher et al. We found that in none of the samples did the larger half give

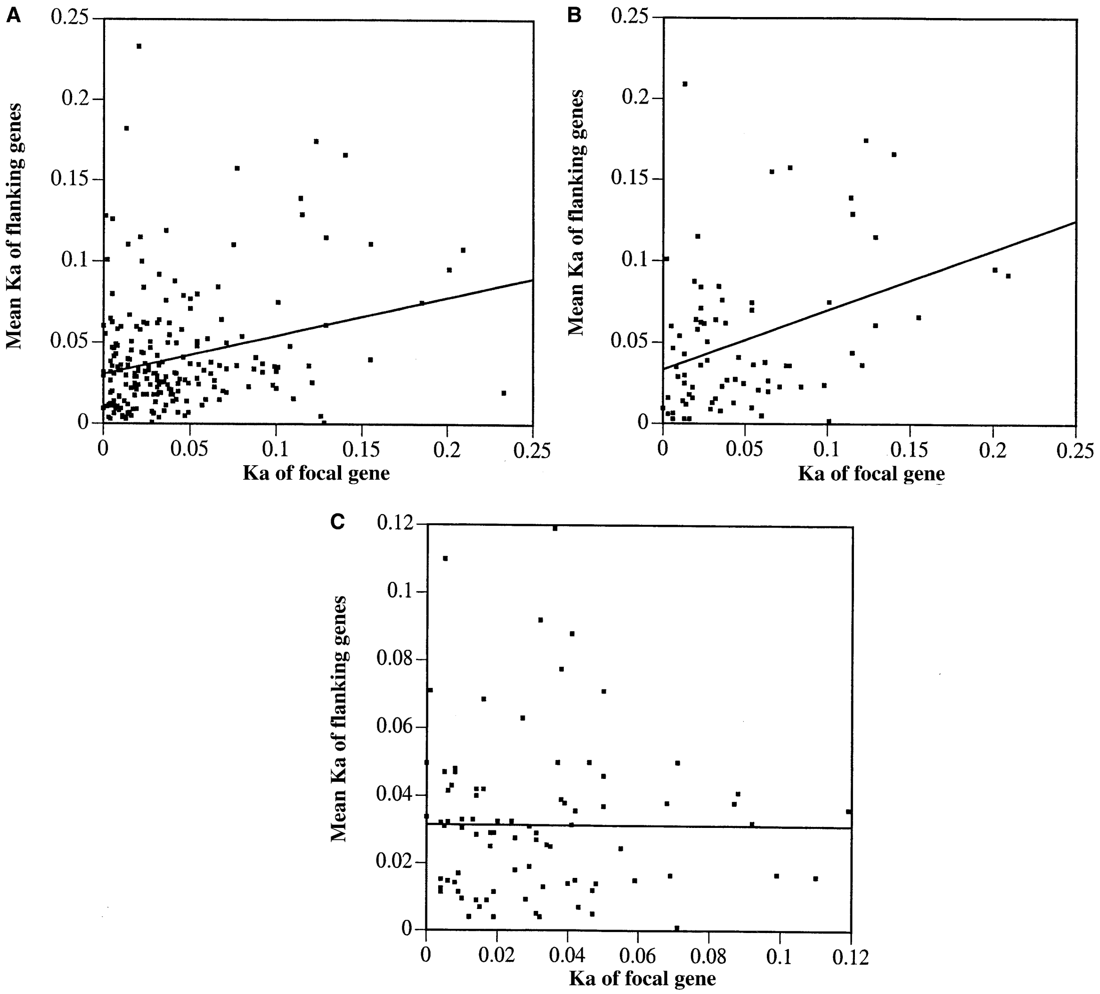

Fig. 2. The relationship between the K_a of a focal gene and the mean K_a of the surrounding genes for **(A)** the complete sample, **(B)** the tissue-specific genes, and **(C)** the non-tissue-specific genes.

an r^2 near 0.0% or the small half give such a high r^2 value. This indicates that this result is not an artifact of subsampling per se. As regards K_a/K_s, the r^2 in the non-tissue-specific sample is approximately half that in the complete data set. In none of 100 trials did this amount of diminution occur.

What classes of tissue-specific genes are there that could possibly be responsible for this effect? There is an a priori expectation that genes involved in antagonistic coevolution may evolve at high rates. If these too are clustered, then this will lead to local similarity in the rates of evolution. This is because in randomized data sets on the average, these fast-evolving genes tend to be paired up with slower-evolving genes. They would thus cause the randomized data sets to have a higher average difference in K_a between linked genes than in the real data set. We have shown that immune genes tend to cluster and it is well established that they have unusually high rates of evolution, probably because of host parasite coevolution. Similarly, genes putatively involved in mother–offspring conflict may show increased rates of evolution (Hurst and McVean 1998). Many of these are likely to be placentally expressed. This is of significance, as prior evidence suggests that placentally expressed genes are clustered (Ko et al. 1998). We therefore examined the consequences of removal of immune and placental genes. Given the absence of a priori expectations for other sets of genes for which we have data, we shall not examine any other subcategories.

Removal of the immune system genes has a little effect on the extent of local similarity as assayed by the r^2 values. Now using 243 genes (i.e., the complete set minus the immune genes), we find a comparable amount of local similarity in K_a as in the complete data set (K_a, $r^2 = 7.9\%$ and $p = 0.05$; K_a/K_s, $r^2 = 8.1\%$ and $p = 0.12$) (Table 1a). When we examine 100 random data sets, each containing 243 randomly selected entries for the original data set, we find that the r^2 value from the nonimmune data set is not unusual. Indeed, in the case of K_a, the r^2 increases. This indicates that the clustering of immune genes is not of importance in determining the local similarity in rates of evolution.

Given the lack of effect on the r^2 values, the decline in the p value seen on the removal of immune genes most likely reflects sample size changes. This we confirmed. We took each of the 100 randomly assembled data sets of 243 genes and measured the mean local similarity within each using the method of Lercher et al. Then we did 10,000 randomizations of each of these 100. We then asked what proportion showed a greater mean local similarity and thereby determined a p value for each of the 100 sets. We found that 12% of the random collections reported a p value above that shown in the nonimmune data set. We therefore failed to reject the hypothesis that the weakening of the p value in the nonimmune set is anything other than a sampling effect.

These conclusions are further supported by analysis of the tissue-specific genes. Within the tissue-specific group without the immune genes, the local similarity is increased (from $r^2 = 9.6\%$ in the complete set of tissue-specific genes to $r^2 = 20.5\%$ after the removal of immune genes) under the protocol of Williams and Hurst (2000). Under the protocol of Lercher et al (2001), the r^2 remains largely unchanged ($r^2 = 14$ versus 13.3%).

The distribution of seven placentally expressed genes, in contrast, appears to have an effect on the local similarity within the class of tissue-specific genes. When these are removed from the tissue-specific gene class, the local similarity decreases and is not statistically significant under either model (method of Lercher et al.—K_a, $p = 0.125$ and $r^2 = 9.5\%$; method of Williams and Hurst—K_a, $p = 0.078$ and $r^2 = 5.0\%$). Again using the method of randomly subsampling, this time randomly removing seven genes from the tissue-specific data set, none of the 100 random subsamples showed such dramatic decreases in r^2.

Discussion

In this paper we have set out to ask two questions. First, do similarly expressed genes tend to cluster in the genome? Second, if they do, does this explain why linked genes evolve at similar rates? We have found evidence that there is a significant tendency for genes of comparable expression breadth to be linked but only a weak tendency, at most, for genes that are coexpressed to be linked. One limitation of our study is the usage of expression data that permit us to analyze presence or absence of expression rather than rate of expression, which might be the more relevant parameter. In the near-future results from microarray data and SAGE analyses should allow exploration of these issues as well.

Given the weakness of the tendency for genes of comparable expression to be linked, and the weakness of the correlation between K_a and expression breadth, we might reasonably conclude that it is a priori unlikely that linkage of similarly expressed genes might explain why linked genes evolve at similar rates. This, however, appears not to be so: within the class of nonspecific genes there is no tendency for linked genes to have similar rates of protein evolution. The local similarity in rates of evolution appears to be due in no small part to the genomic positioning of tissue-specific genes. This is due in part to clustering of placentally expressed genes but is not dependent on the clustering of immune system genes.

These results do not examine whether coexpression more generally underlies local similarity in rates of evolution. Unfortunately, here we can perform only much weaker tests. In yeast, the member of a pair of duplicates that has the higher expression level has the higher rate of protein evolution (Pal et al. 2001b). Evidence that this is so came from analysis of the regression of the difference in the rate of protein evolution versus the difference in expression level (assayed by microarray data) for each pair of duplicates. We can attempt the same sort of analysis for the present data set. That is, if similarity of expression pattern does explain some of the local similarity of rates of protein evolution, then we expect that a large local difference in K_a should reflect a large difference in expression profile.

To see whether this occurs we can plot ΔK_a (pairwise difference in K_a) versus ICE for each pair of linked genes. If coexpression predicts the local similarity in K_a to any extent, then we expect a negative correlation between ΔK_a and ICE. We do not find this: ΔK_a versus ICE, Model 1 ($\Delta K_a = 0.03 - 0.001$ ICE1; $r^2 = 0.001\%$, $p = 0.61$); ICE, Model 2 ($\Delta K_a = 0.03 - 0.005$ ICE2; $r^2 = 0.006\%$, $p = 0.295$; and ICE, Model 3 ($\Delta K_a = 0.03 - 0.006$ ICE3; $r^2 = 0.009\%$, $p = 0.175$). However, while we know that the expression breadth covaries with K_a, ΔK_a does not covary with ΔE ($\Delta K_a = 0.0324 - 0.0001\ \Delta E$; $r^2 = 0.00\%$, $p = 0.8$). The latter result indicates that these are very weak tests. The above result must therefore be considered a rejection of the possibility that there is a strong covariation of expression and rate of evolution. We cannot therefore make any strong conclusions regarding coexpression.

The GC K$_a$ *Problem*

It is remarkable that removal of the tissue-specific genes from the data set destroys the signal of local similarity in rates of protein evolution. This suggests that the effects are unlikely to be genome-wide. This, however, leaves the problem of the causes of the negative correlation between GC content and K_a. Unlike the K_s/GC and K_a/K_s correlations, the K_a/GC correlation is not sensitive to method: the GY94 protocol reports the same result as Li93 (Li93, $K_a = 0.108218 - 0.118808$GC4, $r^2 = 13.1\%$, $p < 0.0001$; GY94, K_a_ML $= 0.122907 - 0.141966$GC4, $r^2 = 16.4\%$, $p < 0.0001$). This negative correlation was considered by Williams and Hurst (2000) to be consistent with the idea that local similarity in rates of protein evolution was due to genome-wide variation in the strength of purifying selection owing to variation in the recombination rate around the genome (i.e., the intensity of Hill–Robertson effects). This interpretation rests on the understanding that the recombination rate covaries with the GC content (Fullerton et al. 2001). The Hill–Robertson model is given some support by the finding that variation in K_a and K_a/K_s within the *Drosophila* genome covaries negatively with the recombination rate (Comeron and Kreitman 2000).

If the local similarity in rates of protein evolution is due largely to linkage of similarly expressed genes, and disappears when tissue-specific genes are removed, how are we to interpret this strong GC/K_a correlation? One possibility is that, as in yeast, the recombination rate (hence GC) and gene expression rates covary, so a correlation between recombination/GC and K_a need not be evidence for Hill–Robertson effects (Pal et al. 2001a). We cannot analyze expression rates in mammals. We find, however, that there is a positive correlation between breadth of expression and GC content at fourfold redundant sites ($E = 4.28 + 4.27\text{GC4}$, $r^2 = 1.3\%$, $p = 0.06$). Given that broadly expressed genes have low rates of evolution, this is in the right direction to explain why genes with a high GC content might have low rates of protein evolution. The correlation is, however, weak [and, incidentally, in the direction opposite to that reported by Goncalves et al. (2000) for human sequences]. This effect is so weak that it cannot account for the greatly reduced K_a in regions of high GC content. This is confirmed by the finding that the K_a/GC correlation remains when only the tissue-specific genes are analyzed ($K_a = 0.15 - 0.18\text{GC4}$, $r^2 = 22.1\%$, $p < 0.0001$, $N = 126$).

Alternatively, it might simply be the case that immune system genes (under directional selection or subject to overdominance) tend to be AT rich. Were this so, the GC/K_a correlation need not be indicative of variation in purifying selection. Indeed, when we divided our data set into immune and nonimmune genes, immune system genes tended to be AT rich (GC4 immune = 0.55 ± 0.016; GC4 nonimmune = 0.61 ± 0.008). However, both sets still showed a strong K_a/GC4 correlation (nonimmune, $K_a = 0.075 - 0.075\text{GC4}$, $r^2 = 8.6\%$, $p < 0.001$; immune, $K_a = 0.21 - 0.24\text{GC4}$, $r^2 = 24.1\%$, $p = 0.001$).

Given that these two possible explanations do not explain the GC/K_a correlation, we must regard the cause as problematic. Given that the result is both relatively strong and robust to methodology (unlike the K_a/K_s correlation), the causes of the correlation deserve further scrutiny.

Acknowledgments. We thank Deborah Charlesworth, Martin Lercher, and Araxi Urratia for their constructive comments.

References

Bortoluzzi S, Rampoldi L, Simionati B, Zimbello R, Barbon A, d'Alessi F, Tiso N, Pallavicini A, Toppo S, Cannata N, Valle G, Lanfranchi C, Danieli GA (1998) A comprehensive, high-resolution genomic transcript map of human skeletal muscle. Genome Res 8:817–825

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A, Versteeg R (2001) The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. Science 291:1289–1292

Comeron JM, Kreitman M (2000) The correlation between intron length and Recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. Genetics 156:1175–1190

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

Duret L, Mouchiroud D, Gouy M (1994) Hovergen—A database of homologous vertebrate genes. Nucleic Acids Res 22:2360–2365

Fullerton SM, Carvalho AB, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18:1139–1142

Goldman N, Yang ZH (1994) Codon-based model of nucleotide substitution for protein-coding dna sequences. Mol Biol Evol 11:725–736

Goncalves I, Duret L, Mouchiroud D (2000) Nature and structure of human genes that generate retropseudogenes. Genome Res 10:672–678

Hughes AL, Yeager M (1997) Molecular evolution of the vertebrate immune system. Bioessays 19:777–786

Hurst LD, McVean GT (1998) Do we understand the evolution of genomic imprinting? Curr Opin Genet Dev 8:701–708

Hurst LD, Smith NGC (1999) Do essential genes evolve slowly? Curr Biol 9:747–750

Ko MSH, Threat TA, Wang XQ, Horton JH, Cui YS, Wang XH, Pryor E, Paris J, WellsSmith J, Kitchen JR, Rowe LB, Eppig J, Satoh T, Brant L, Fujiwara H, Yotsumoto S, Nakashima H (1998) Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. Hum Mol Genet 7: 1967–1978

Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lercher MJ, Williams EJB, Hurst LD (2001) Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons. Mol Biol Evol 18:2032–2039

Li W-H (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

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

Pal C, Bapp B, Hurst LD (2001b) Highly expressed genes in yeast evolve slowly. Genetics 158:927–931

Pamilo P, Bianchi NO (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol 10:271–281

Smith NGC, Hurst LD (1999) The causes of synonymous rate variation in the rodent genome: Can substitution rates be used to estimate the sex bias in mutation rate? Genetics 152:661–673

Williams EJB, Hurst LD (2000) The proteins of linked genes evolve at similar rates. Nature 407:900–903